ural waves is much smaller than the overall curvature, the vibrational modes will be closely related to the flexural waves of a circular disc, with frequencies satisfying Chladni's generalised empirical law, $f_{mn} \sim c(m+2n)^p$, as confirmed by *Perrin* et al. [15.218].

The flexural modes involve radial displacements proportional to $\cos(m\phi)$. Continuity requires that there must also be a tangential displacement, such that $u + \partial v/\partial\phi = 0$, where u and v are the radial and tangential velocities respectively. Coupling to the tangential motion explains why it is possible to feed energy into a vibrating wine glass or the individual glass resonators of a glass harmonica (*Rossing* [15.201] Chap. 14), by rubbing a wetted finger around the rim. The excitation is very similar to the slip–stick mechanism used to excite the bowed string (Sect. 15.2.4).

Figure 15.136 illustrates a set of holographic measurements by *Rossing* ([15.201], Fig. 12.4), which is typical of most bell shapes. The $(m, 1)$ and $(m, 2)$ modes can immediatly be related to the (m, n) modes of a cupped disc. However, there is a distinct change in character for $n > 0$, with an additional node appearing close to the rim – referred to as a $(m, 1\#)$ mode. Some insight is provided by the finite element solutions for a typical large English church bell illustrated in Fig. 15.137 [15.221]. The three modes illustrated are very similar to the $(3, 0)$, $(3, 1)$ and $(3, 2)$ modes expected from the simple cupped disc model, except for the lowest frequency $(3, 0)$ mode,

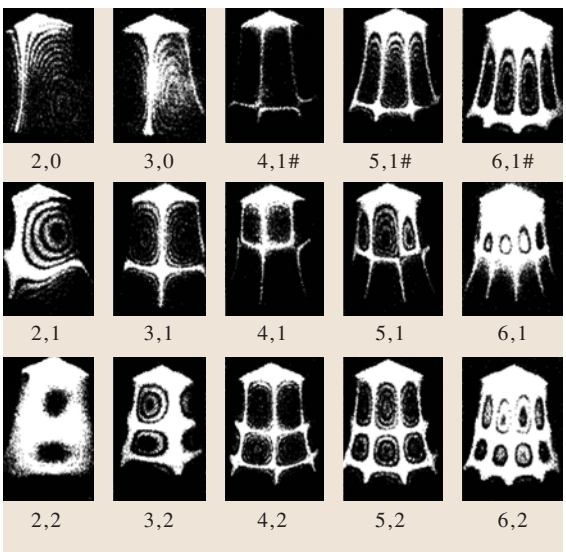


Fig. 15.136 Holographic interferograms and nomenclature for vibrational modes of a hand bell (after *Rossing* [15.201])

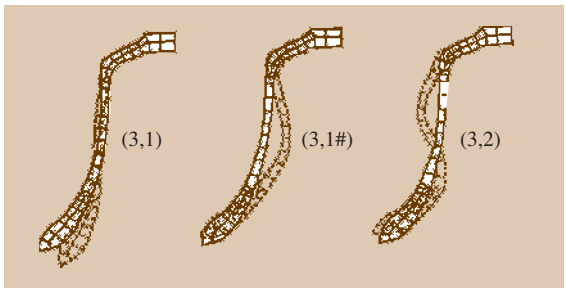


Fig. 15.137 Finite-element solutions for the lowest-order $m = 3$ modes of an English church bell (after *Perrin* et al. [15.219]) illustrating the (3,1), (3,1#) and (3,2) modes

in which the top surfaces of the bell move in antiphase with the rim, to give a nodal line about half-way along the length. Similarly, the anticipated $(4, 0)$, $(5, 0)$ and $(6, 0)$ modes of the handbell investigated by *Rossing* [15.201] acquire an additional nodal line close to the end rim denoted as $(4, 1\#)$, $(5, 1\#)$ and $(6, \#)$ modes. Such features can generally only be accounted for by detailed computational analysis.

Bell Tuning

Figure 15.138 shows the frequencies and associated modes of a well-tuned traditional church bell (*Rossing* and *Perrin* [15.220]), ordered into groups based on mode

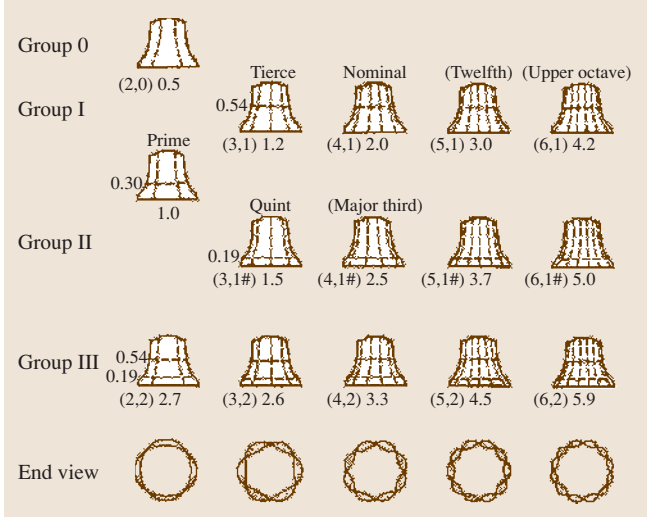


Fig. 15.138 Measured frequencies for a typical D5 church bell, indicating the relative frequencies of the observed mode, with the traditional names associated with such modes indicated (after *Rossing* and *Perrin* [15.220])

shapes, which enable correlations to be made between bells of different shapes and sizes.

Fine-quality bells are carefully tuned by the bell maker, so that the lowest modes have harmonically related frequencies, thereby achieving a well-defined sense of pitch. This is achieved by selective thinning of the thickness of the bell on a very large lathe after its initial casting. In an ideally tuned bell, the principal modes are designated as the hum (2,0) mode, an octave below the prime or fundamental (2,1#) mode. A minor third above the prime (ratio 5:4) is the tierce (3,1) mode, a perfect fifth above that (ratio 3:2) is the quint or fifth (3,1#) mode, and an octave above is the nominal (4,1) mode. *Fletcher* and *Rossing* ([15.5], Table 21.1) compare these and higher modes with measured values for a particular bell.

The art of tuning the partials of bells to achieve a well-sounding note was initially developed in the seventeenth century in the low countries (what is now Holland and the northern parts of Germany and France). Many fine bells from this period by François and Pieter Hemony are still in use in carillons today (*Rossing* [15.201]). However, the art of bell tuning then appears to have been lost until the end of the 19th century, when it was rediscovered by Canon Arthur Simpson in England, following Lord Rayleigh's pioneering research on the sound of bells ([15.3] Vol. 1, Sect. 235).

Acoustic Radiation

The strongest-sounding partials of most bells are the group I $(m,1)$ modes, with a nodal circle approximately halfway up the bell (*Fletcher* and *Rossing* [15.5], Sect. 21.11). Such modes have $2m$ antinodal areas providing spatially alternating sound sources in antiphase. The radiation efficiency of such modes increases rapidly with size of bell and frequency. If the acoustic wavelength is much larger than the separation of such antinodes, the sound from such sources will tend to cancel out. However, above a crossover frequency, such that the velocity of sound is equal to that of the flexural waves on the surface of the bell, $v_{\text{flex}} = \sqrt{1.8 c_{\text{L}} h f}$, the spacing between the antinodes will exceed the wavelength in air and the modes will radiate more efficiently. For large church bells, this condition is satisfied for almost all but the very lowest partials, so almost all partials radiate sound rather efficiently. In contrast, hand bells with rather thin walls are significantly less efficient. There is also a small intensity of sound radiated axially at double the modal frequencies, from the induced fluctuations in the volume of air enclosed within the bell.

When a bell is struck, usually by a cast or wrought-iron, ball-shaped, clapper, the first sound heard, the strike note, contains contributions from the very large range of largely inharmonic partials of the bell. Nevertheless the listener can usually identify a pitch to the initial note, determined by the prominent partials excited. However, it is not always easy to attribute the pitch of a note to a particular octave, which is a common feature of many struck percussion instruments (e.g. notes on a xylophone, steeldrum and even timpani). The pitch of the strike note appears to be determined principally by the excited partials with frequencies in the ratios 2:3:4. The ear attributes the pitch to be that of the missing fundamental an octave below the lowest of the partials principally excited, which does not necessarily correspond to the pitch of the lowest partial excited, unless the bell is well tuned.

The majority of the higher partials decay rather quickly, with the long-term sound dominated by the hum note. For a 70 cm church bell, *Perrin* et al. [15.218] measured T_{60} decay times of 52 s for the (2,0) hum mode, 16 s for the (2.1#) and (3,1) prime and minor third modes, 6 s for the (4,1) octave and 3 s for the (4,1#) upper major third, with progressively shorter decay times

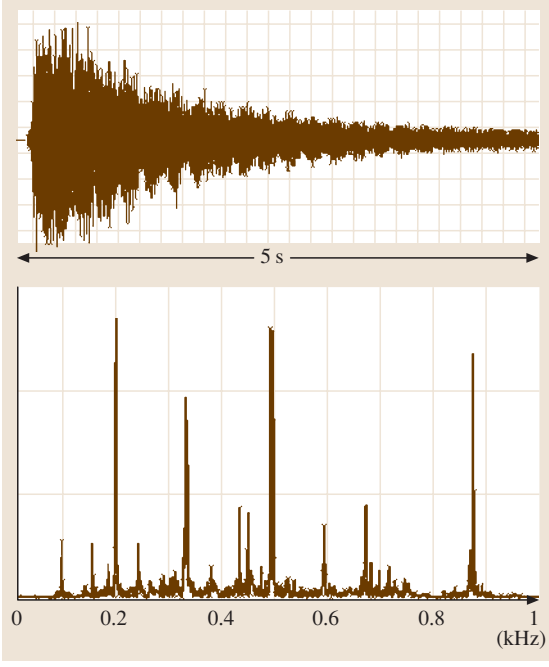


Fig. 15.139 Decaying vibrations and spectrum of Big Ben, London

for the higher modes. Audio CD-ROM illustrates the synthesised sound of the above harmonic modes excited with equal amplitudes. This is followed by a synthesised bell with a major rather than a minor third partial. Such a bell has recently been realised (1999) based on FEA studies at the Technical University in Eindhoven in collaboration with the Eijsbouts Bell foundry. It produces a less mournful sound when played in the major scale music in a carillon. In both cases the synthesised sound reproduces the gentle sound of a bell rather well, though it lacks the initial clang that results from the higher-frequency inharmonic modes of a real bell.

As an example of the waveform and spectrum of a large bell, in Fig. 15.139 and audio signal CD-ROM, we illustrate the sound and rich spectrum of partials of one of the most famous large bells in the world, Big Ben, hung high above the Houses of Parliament in London. This bell is broadcast each day following the Westminster Chimes, marking the hours and end of broadcasting on the BBC each night for UK listeners.

Non-Axial Symmetry

Bell modes with $m > 1$ are doubly degenerate with orthogonal modes varying as $\cos m\phi$ and $\sin m\phi$ and nodal lines in the azimuthal directions that bisect each other. Any departures from axial symmetry will lift the degeneracy and give rise to a split pair of orthogonal modes. If both modes are excited together, the two frequencies will beat against each other, as already evident in the sound of Big Ben and in the sound (CD-ROM) of a slightly asymmetrical glass beaker with a pouring spout, which lifts the degeneracy of the otherwise axially symmetric modes.

Bells with strongly distorted or elliptical cross sections can have two completely different pitches, which can be sounded independently by beating at the antinodal positions of one set of modes mode and nodes of the other. Drums with quasi-elliptical cross sections will therefore sound a single note when struck at the narrowest and widest cross-sectional radii, and a second note, when struck at appropriate positions in between.

A dramatic example of axially asymmetric bells is provided by the 65 ancient bells from the tomb of Zeng Hou Yi found at Sui Xiang from around 433 BC. These bells from the second millennium BC are masterpieces of Chinese art and bell casting. They have oval cross sections and range from small hand bells to well over 1.5 m in height. When struck at different positions along the flattened surfaces, two quite distinct tones can be excited, which were designed to be about a major or minor third apart. For further details of these and other Chinese and other eastern bells see *Rossing* ([15.201] Chap. 13) and [15.222].

References

15.1 T. Levenson: *Measure for Measure: How Music and Science together have explored the Universe* (Oxford Univ. Press, Oxford 1997)

15.2 S. Hawkins (Ed.): *On the Shoulders of Giants* (Running, Philadelphia 2002)

15.3 Lord Rayleigh: *The Theory of Sound: I and II (1896)*, 2nd edn. (Dover, New York 1945), Reprint

15.4 S. Hawkins: *The Universe in a Nutshell* (Bantam, London 2001)

15.5 N.H. Fletcher, T.D. Rossing: *The Physics of Musical Instruments*, 2nd edn. (Springer, New York, Berlin 1998)

15.6 T.D. Rossing, F.R. Moore, P.A. Wheeler: *The Science of Sound*, 3rd edn. (Addison Wesley, San Francisco 2002)

15.7 D.M. Campbell, A. Myers, C.A. Greated: *Musical Instruments: History, Technology and Performance of Instruments of Western Music* (Oxford Univ. Press, Oxford 2006)

15.8 G. Bissinger: Contemporary generalised normal mode violin acoustics, Acustica **90**, 590–599 (2004)

15.9 A. Hirschberg, J. Gilbert, R. Msallam, A.P.J. Wijnands: Shock waves in trombones, J. Acoust. Soc. Am. **99**, 1754–1758 (1996)

15.10 T.J.W. Hill, B.E. Richardson, S.J. Richardson: *Modal Radiation from Classical Guitars: Experimental Measurements and Theoretical Predictions. Proc SMAC 03* (KTH, Stockholm 2003) pp. 129–132

15.11 L. Cremer: *The Physics of the Violin, section 11.2* (MIT Press, Mass 1984)

15.12 B. Britten: *Serenade for Tenor, Horn and Strings* (Hawkes & Son, London 1944)

15.13 J.M. Barbour: *Tuning and Temperament* (Michigan State College Press, East Lansing 1953)

15.14 National Instruments: *Windowing: Optimizing FFTs Using Windowing Functions* (National Instruments, Austin 2006), http://zone.ni.com/devzone/conceptd.nsf/

15.15 M.-P. Verge, A. Hirschberg: Turbulence noise in flue instruments. In: *Proc. Int. Symp. Mus. Acoust.*, SMAC 95 (IRCAM, Paris 1995) pp. 94–99

15.16 M.E. McIntyre, R.T. Schumacher, J. Woodhouse: Aperiodicity in bowed-string motion: On the dif-

ferential slipping mechanism, Acustica **49**, 13–32 (1981)
15.17 J. Meyer: Zur klanglichen wirkung des streicher-vibratos, Acustica **76**, 283–291 (1992)
15.18 C.E. Gough: Measurement, modelling and synthesis of violin vibrato sounds, Acta Acustica **91**, 229–240 (2005)
15.19 E.I. Prame: Measurement of the vibrato rate of ten singers, J. Acoust. Soc. Am. **96**, 1979–1984 (1994)
15.20 J. Gilbert, L. Simon, J. Terroir: *Vibrato in single reed instruments, SMAC 03* (Royal Swedish Academy of Music, Stockholm 2003) pp. 271–273
15.21 D.W. Robinson, R.S. Dadson: A re-determination of the equal-loudness relations for pure tones, Brit. J. Appl. Phys. **7**, 166–181 (1956)
15.22 H.A. Fletcher, W.A. Munson: Loudness, its definition, measurement and calculation, J. Acoust. Soc. Am. **9**, 82–108 (1933)
15.23 B. Patterson: Musical dynamics, Sci. Am. **231**(5), 78–95 (1974)
15.24 B.C.J. Moore: *Psychology of Hearing* (Academic, London 1997)
15.25 A.J.M. Houtsma, T.D. Rossing, W.M. Wagenaars: *Auditory Demonstrations, Phillips compact disc 1126-061* (Acoust. Soc. Am., Sewickley 1989)
15.26 C.M. Hutchins (Ed.): *Musical Acoustics, Part I (Violin Family Components) and II (Violin Family Functions) Benchmark papers in Acoustics, Nos. 5 and 6* (Dowden Hutchinson, Ross, Stroudsburg 1975,1976)
15.27 C. Hutchins, V. Benade (Eds.): *Research Papers in Violin Acoustics 1975-1993, Vols. 1 and 2*, (Acoust. Soc. Am., Melville 1997)
15.28 CAS Newsletters (1994-1987), CAS Journals (1988-2004), www.catgutacoustical.org
15.29 L. Cremer: *The Physics of the Violin* (MIT Press, Cambridge 1984)
15.30 R. Midgley: *Musical Instruments of the World* (Paddington, New York 1978)
15.31 A. Baines: *European and American Musical Instruments* (Chancellor, London 1983)
15.32 G.G. Stokes: *On the communication of vibrations from a vibrating body to a surrounding gas* (Philos. Trans., London 1868)
15.33 C.E. Gough: The resonant response of a violin G-string and the excitation of the wolf-note, Acustica **44**(2), 673–684 (1980)
15.34 H.F. Helmholtz: *The Sensation of Tone (1877), translated by A.J. Ellis* (Dover, New York 1954)
15.35 C.V. Raman: On the mechanical theory of the vibrations of bowed strings and of musical instruments of then violin family, with experimental verification of the results: Part 1, Indian Assoc. Cultivation Sci. Bull. **15**, 1–158 (1918)
15.36 J.C. Schelling: The bowed string and player, J. Acoust. Soc. Am. **9**, 91–98 (1973)
15.37 S. Thwaites, N.H. Fletcher: Some notes on the clavichord, J. Acoust. Soc. Am. **69**, 1476–1483 (1981)
15.38 D.E. Hall: Piano string excitation in the case of a small hammer mass, J. Acoust. Soc. Am. **79**, 141–147 (1986)
15.39 D.E. Hall: Piano string excitation II: General solution for a hard narrow hammer, J. Acoust. Soc. Am. **81**, 535–546 (1987)
15.40 D.E. Hall: Piano string excitation III: General solution for a soft narrow hammer, J. Acoust. Soc. Am. **81**, 547–555 (1987)
15.41 D.E. Hall: Piano string excitation IV: Non-linear modelling, J. Acoust. Soc. Am. **92**, 95–105 (1992)
15.42 P.M. Morse, K.U. Ingard: *Theoretical Acoustics* (McGraw-Hill, New York 1968), Reprinted by Princeton Univ. Press, Princeton 1986
15.43 H. Fletcher: Normal vibration frequencies of a stiff piano string, J. Acoust. Soc. Am. **36**, 203–209 (1964)
15.44 E.L. Kent: Influence of irregular patterns in the inharmonicity of piano-tone partials upon tuning practice, Das Musikinstrument **31**, 1008–1013 (1982),
15.45 N.C. Pickering: *The Bowed String* (Bowed Instruments, Southampton 1991)
15.46 N.C. Pickering: Physical Properties of Violin Strings, J. Catgut Soc. **44**, 6–8 (1985), paper 21 in ref [15.42]
15.47 C. Vallette: The mechanics of vibrating strings. In: *Mechanics of musical instruments*, ed. by A. Hirschberg, J. Kergomard, G. Weinreich (Springer, Berlin, New York 1995) pp. 115–183
15.48 K. Legge, N.H. Fletcher: Nonlinear generation of missing modes on a vibrating string, J. Acoust. Soc. Am. **76**, 5–12 (1984)
15.49 C.E. Gough: The nonlinear free vibration of a damped elastic string, J. Acoust. Soc. Am. **75**, 1770–1776 (1984)
15.50 J. Miles: Resonant non-planar motion of a stretched string, J. Acoust. Soc. Am. **75**, 1505–1510 (1984)
15.51 R.J. Hanson, J.M. Anderson, H.K. Macomber: Measurement of nonlinear effects in a driven vibrating wire, J. Acoust. Soc. Am. **96**, 1549–1556 (1994)
15.52 R.J. Hanson, H.K. Macomber, A.C. Morrison, M.A. Boucher: Primarily nonlinear effects observed in a driven asymmetrical vibrating wire, J. Acoust. Soc. Am. **117**, 400–412 (2005)
15.53 J.A. Elliot: Intrinsic nonlinear effects in vibrating strings, Am. J. Phys. **48**, 478–480 (1980)
15.54 J. Woodhouse, P.M. Galluzzo: The bowed string as we know it today, Acustica **90**, 579–590 (2004)
15.55 F.A. Saunders: Recent work on violins, J. Acoust. Soc. Am. **25**, 491–498 (1953)
15.56 J.C. Schelling: The bowed string and the player, J. Acoust. Soc. Am. **53**, 26–41 (1973)
15.57 J.C. Schelling: The physics of the bowed string. In: *The Physics of Music*, ed. by C. Hutchins, Scientific American (W.H. Freeman & Co., San Fransisco 1978)
15.58 R.T. Schumacher: Measurements of some parameters of bowing, J. Acoust. Soc. Am. **96**, 1985–1998 (1994)

Amazon EX1014 Page 671

15.59 F.G. Friedlander: On the oscillation of the bowed string, Proc. Cambridge Phil. Soc. **49**, 516–530 (1953)

15.60 M.E. McIntyre, J. Woodhouse: On the fundamentals of bowed string dynamics, Acustica **43**, 93–108 (1979)

15.61 J. Woodhouse: On the playability of violins, Part II: Minimum bow force and transients, Acustica **78**, 137–153 (1993)

15.62 M.E. McIntyre, R.T. Schumacher, J. Woodhouse: Aperiodicity in bowed-string motion: On the differential slipping mechanism, Acustica **49**, 13–32 (1981)

15.63 R.T. Schumacher: Oscillations of bowed strings, J. Acoust. Soc. Am. **43**, 109–120 (1979)

15.64 K. Guettler: On the creation of the Helmholtz motion in bowed strings, Acustica **88**, 2002 (2002)

15.65 P.M. Galluzzo: *On the Playability of stringed instruments, PhD Thesis* (Cambridge University, Cambridge 2003)

15.66 J.H. Smith, J. Woodhouse: The tribology of rosin, J. Mech. Phys. Solids **48**, 1633–1681 (2000)

15.67 W. Reinicke: *Dissertation, Institute for Technical Acoustics* (Technical University of Berlin, Berlin 1973)

15.68 J. Woodhouse, R.T. Schumacher, S. Garoff: Reconstruction of bowing point friction force in a bowed string, J. Acoust. Soc. Am. **108**, 357–368 (2000)

15.69 J.H. Smith: *Stick-slip vibration and its constitutive laws, PhD thesis* (Cambridge University, Cambridge 1990)

15.70 J. Woodhouse: Bowed string simulation using a thermal friction model, Acustica **89**, 355–368 (2003)

15.71 W. Reinicke, L. Cremer: Application of holographic interferometry to the bodies of sting instruments, J. Acoust. Soc. Am. **48**, 988–992 (1970)

15.72 J. Woodhouse: data kindly provided for this chapter

15.73 H. Dünnewald: Ein erweitertes Verfahren zur objektiven Bestimmung der Klangqualität von Violinen, Acustica **71**, 269–276 (1990), reprinted in English as Deduction of objective quality parameters on old and new violins, J. Catgut Acoust. Soc. 2nd Ser. 1(7) 1-5 (1991), included in Hutchins [15.27, Vol 1]

15.74 E.V. Jansson: Admittance measurements of 25 high quality violins, Acta Acustica **83**, 337–341 (1997)

15.75 E.V. Jansson: Violin frequency response – bridge mobility and bridge feet distance, Appl. Acoust. **65**, 1197–1205 (2004)

15.76 J. Woodhouse: On the "bridge hill" of the violin, Acustica **91**, 155–165 (2005)

15.77 J.A. Moral, E.V. Jansson: Eigenmodes, input admittance and the function of the violin, Acustica **50**, 329–337 (1982)

15.78 M. Hacklinger: Violin timbre and bridge frequency response, Acustica **39**, 324–330 (1978)

15.79 D. Gill: *The Book of the Violin*, ed. by D. Gill (Phaedon, Oxford 1984) p. 11, Prologue

15.80 C.E. Gough: The theory of string resonances on musical instruments, Acustica **49**, 124–141 (1981)

15.81 J. Woodhouse: On the synthesis of guitar plucks, Acustica **89**, 928–944 (2003)

15.82 J. Woodhouse: Plucked guitar transients: comparison of measurements and synthesis, Acustica **89**, 945–965 (2003)

15.83 G. Weinreich: Coupled piano strings, J. Acoust. Soc. Am. **62**, 1474–1484 (1977)

15.84 C.G.B. Baker, C.M. Thair, C.E. Gough: A photodetector for measuring resonances of violin strings, Acustica **44**, 70 (1980)

15.85 F. Savart: *Mémoires sur le construction des Instruments Ã Cordes et Ã Archet* (Deterville, Paris 1819), See also the brief summary and references to modern translations of Savart's research in Hutchins 15.26 Part 1, page 8

15.86 M.E. McIntyre, J. Woodhouse: On measuring the elastic and damping of orthotropic sheet materials, Acta. Metall. **36**, 1397–1416 (1988)

15.87 M.D. Waller: Vibrations of free rectangular plates, Proc. Phys. Soc. London B **62**, 277–285 (1949)

15.88 A.W. Leissa: *Vibration of Plates* (NASA SP-160, Washington 1993), Reprinted by Acoust. Soc. Am., Woodbury 1993

15.89 C. Hutchins: The Acoustics of Violin Plates, Sci. Am. **1**, 71–186 (1981)

15.90 B.E. Richardson, G.W. Roberts: *The adjustment of mode frequencies in guitars: a study by means of holographic interferometry and finite element analysis. SMAC 83* (Royal Swedish Academy of Music, Stockholm 1985) pp. 285–302

15.91 C. Hutchins: A rationale for BI-TRI octave plate tuning, Catgut Acoust. Soc. J. **1**(8), 36–39 (1991), Series II

15.92 E. Reissner: On axi-symmetrical vibrations of shallow spherical shells, Q. Appl. Math. **13**, 279–290 (1955)

15.93 K.D. Marshall: Modal analysis of a violin, J. Acoust. Soc. Am. **77**(2), 695–709 (1985)

15.94 G. Bissinger: Contemporary generalised normal mode violin acoustics, Acustica **90**, 590–599 (2004)

15.95 N.J.-J. Fang, O.E. Rodgers: Violin soundpost elastic vibration, Catgut Acoust. Soc. J. **2**(1), 39–40 (1992)

15.96 G.A. Knott: *A modal Analysis of the Violin using MSC.NASTRAM and PATRAN. MSc Thesis* (Naval Postgraduate School, Monterey 1987), Reproduced in Hutchins/Benade 15.27

15.97 J.P. Beldie: *Dissertation* (Technical University of Berlin, Berlin 1975), as described in Cremer 15.29 Sect. 10.4-5

15.98 J. Meyer: Die Abstimmung der Grundresonanzen von Guitarren, Das Musikinstrument **23**, 179–186

Amazon EX1014 Page 672

(1974), English translation in J. Guitar Acoustics No.5, 19 (1982)
15.99 Q. Christensen: Qualitative models for low frequency guitar function, J. Guitar Acoust. **6**, 10–25 (1982)
15.100 T.D. Rossing, J. Popp, D. Polstein: *Acoustical Response of Guitars. SMAC 83* (Royal Swedish Academy of Music, Stockholm 1985) pp. 311–332
15.101 E.V. Jansson: On higher air modes in the violin, Catgut Acoust. Soc. Newsletter **19**, 13–15 (1973), Reprinted in Musical Acoustics, Part 2, ed. C.M.Hutchins (Dowden, Hutchinson & Ross, Stroudsberg 1976)
15.102 G. Derveaux, A. Chaigne, P. Joly, J. Bécache: Time-domain simulation of guitar: Model and method, J. Acoust. Soc. Am. **114**, 3368–3383 (2003)
15.103 G. Roberts, G. Finite: *Element analysis of the violin, PhD Thesis* (Cardiff University, Cardiff 1986), extract reproduced in Hutchins and Benade 15.27, pp.575–590
15.104 B.E. Richardson, G.W. Roberts: *The adjustment of mode frequencies in guitars: A study by means of holographic and finite element analysis. Proc. SMAC. 83* (Royal Swedish Academy of Music, Stockholm 1985) pp. 285–302
15.105 G. Derveaux, A. Chaigne, P. Joly, J. Bécache: *Numerical simulation of the acoustic guitar* (INRIA, Rocquencourt 2003), from http://www.inria.fr/multimedia/Videotheque-fra.html
15.106 G. Bissinger: A unified materials-normal mode approach to violin acoustics, Acustica **91**, 214–228 (2005)
15.107 M. Schleske: On making "tonal copies" of a violin, J. Catgut Acoust. Soc. **3**(2), 18–28 (1996)
15.108 L.S. Morset: *A low-cost pc-based tool for violin acoustics measurements, ISMA 2001* (Fondazione Scuola Di San Giorgio, Venice 2001) pp. 627–630
15.109 H.O. Saldner, N.-E. Molin, E.V. Jansson: Vibrational modes of the violin forced via the bridge and action of the soundpost, J. Acoust. Soc. Am. **100**, 1168–1177 (1996)
15.110 E.V. Jansson, N.-E. Molin, H. Sundin: Resonances of a violin body studied by hologram interferometry and acoustical methods, Phys. Scipta **2**, 243–256 (1970)
15.111 N.-E. Molin, A.O. Wählin, E.V. Jansson: Transient response of the violin, J. Acoust. Soc. Am. **88**, 2479–2481 (1990)
15.112 N.-E. Molin, A.O. Wählin, E.V. Jansson: Transient response of the violin revisited, J. Acoust. Soc. Am. **90**, 2192–2195 (1991)
15.113 J. Curtin: Innovation in violin making, Proc. Int. Symp. Musical Acoustics, CSA-CAS **1**, 11–16 (1998)
15.114 J. Meyer: Directivity of the bowed string instruments and its effect on orchestral sound in concert halls, J. Acoust. Soc. Am. **51**, 1994–2009 (1972)
15.115 G. Weinreich, E.B. Arnold: Method for measuring acoustic fields, J. Acoust. Soc. Am. **68**, 404–411 (1982)
15.116 G. Weinreich: Sound hole sum rule and dipole moment of the violin, J. Acoust. Soc. Am. **77**, 710–718 (1985)
15.117 T.J.W. Hill, B.E. Richardson, S.J. Richardson: Acoustical parameters for the characterisation of the classical guitar, Acustica **89**, 335–348 (2003)
15.118 G. Bissinger, A. Gregorian: Relating normal mode properties of violins to overall quality signature modes, Catgut Acoust. Soc. J. **4**(8), 37–45 (2003)
15.119 E.V. Jansson: Admittance measurements of 25 high quality violins, Acustica **83**, 337–341 (1997)
15.120 G. Weinreich: Directional tone colour, J. Acoust. Soc. Am. **101**, 2338–2346 (1997)
15.121 J. Meyer: Zur klangichen Wirkung des Streicher-Vibratos, Acustica **76**, 283–291 (1992)
15.122 H. Fletcher, L.C. Sanders: Quality of violin vibrato tones, J. Acoust. Soc. Am. **41**, 1534–1544 (1967)
15.123 M.V. Matthews, K. Kohut: Electronic simulation of violin resonances, J. Acoust. Soc. Am. **53**, 1620–1626 (1973)
15.124 J. Nagyvary: Modern science and the classical violin, Chem. Intel. **2**, 24–31 (1996)
15.125 L. Burckle, H. Grissino-Mayer: Stradivari, violins, tree rings and the Maunder minimum, Dendrichronologia **21**, 41–45 (2003)
15.126 C.M. Hutchins: A 30-year experiment in the acoustical and musical development of violin family instruments, J. Acoust. Soc. Am. **92**, 639–650 (1992),
15.127 H.L.F. Helmholtz: *On the Sensations of Tone*, 4th edn. (Dover, New York 1954), trans. by A.J. Ellis
15.128 H. Bouasse: *Instruments á Vent* (Delagrave, Paris 1929)
15.129 M. Campbell, C. Greated: *The Musicians Guide to Acoustics* (Oxford Univ. Press, Oxford 1987)
15.130 C.J. Nederveen: *Acoustical Aspects of Woodwind Instruments* (Northern Illinois Univ. Press, DeKalb 1998)
15.131 A. Hirschberg, J. Kergomard, G. Weinreich: *Mechanics of Musical Instruments* (Springer, Berlin, New York 1995)
15.132 J. Backus: *The Acoustical Foundations of Music*, 2nd edn. (Norton, New York 1977)
15.133 A.H. Benade: *Fundamentals of Musical Acoustics* (Oxford Univ. Press, Oxford 1975)
15.134 D.M. Campbell (Ed.): Special Issue on Musical Wind Instruments Acoustics, Acustica **86**(4), 599–755 (2000)
15.135 P.M. Moorse, K.U. Ingard: *Theoretical Acoustics* (McGraw-Hill, New York 1968), reprinted by Princeton Univ. Press, Princeton (1986)
15.136 L.L. Beranek: *Acoustics* (McGraw Hill, New York 1954) pp. 91–115, reprinted by Acoust. Soc. Am. (1986)

15.137 H. Levine, L. Schwinger: On the radiation of sound from an unflanged pipe, Phys. Rev. **73**, 383–406 (1948)

15.138 R.D. Ayers, L.J. Eliason, D. Mahrgereth: The conical bore in musical acoustics, Am. J. Phys. **53**, 528–527 (1985)

15.139 H.F. Olson: *Acoustic Engineering* (Van Nostrand-Reinhold, Princeton 1957) pp. 88–123

15.140 L.E. Kinsler, A.R. Frey, A.B. Coppens, J.V. Sanders: *Fundamentals of Acoustics* (Wiley, New York 1982), Fig. 14.19

15.141 A.G. Webster: Acoustical impedance, and the theory of horns and the phonograph, Proc. Nat. Acad. Sci. (US) **5**, 275–282 (1919)

15.142 C.J. Nederveen: *Acoustical Aspects of Woodwind Instruments* (Knuf, Amsterdam 1969) p. 60

15.143 D.H. Keefe, A.H. Benade: Wave propagation in strongly curved ducts, J. Acoust. Soc. Am. **74**, 320–332 (1983)

15.144 R. Caussé, J. Kergomard, X. Luxton: Input impedance of brass instruments – Comparison between experiment and numerical models, J. Acoust. Soc. Am. **75**, 241–254 (1984)

15.145 N.H. Fletcher, R.K. Silk, L.M. Douglas: Acoustic admittance of air-driven reed generators, Acustica **50**, 155–159 (1982)

15.146 A. Almeida, C. Verges, R. Caussé: Quasistatic nonlinear characteristics of double-reed instruments, J. Acoust. Soc. Am. **121**, 536–546 (2007)

15.147 S. Ollivier, J.-P. Dalmont: *Experimental investigation of clarinet reed operation*, *SMAC 03* (Royal Swedish Academy of Music, Stockholm 2003) pp. 283–289

15.148 J.-P. Dalmont, J. Gilbert, S. Ollivier: Nonlinear characteristics of single-reed instruments: Quasistatic flow and reed opening measurements, J. Acoust. Soc. Am. **114**, 2253–2262 (2003)

15.149 J. Backus, C.J. Nederveen: *Acoustical Aspects of Woodwind Instruments* (Knuf, Amsterdam 1969) pp. 28–37

15.150 S. Ollivier, J.-P. Dalmont: *Experimental investigation of clarinet reed operation*, *Proc. SMAC 03* (Royal Swedish Academy of Music, Stockholm 2003) pp. 283–289

15.151 A.P.J. Wijnands, A. Hirschberg: Effect of pipeneck downstream of a double reed. In: *Proc. Int. Symp. Musical Acoustics* (IRCAM, Paris 1995) pp. 148–151

15.152 N.H. Fletcher: Excitation mechanisms in woodwind and brass instruments, Acustica **43**, 63–72 (1979), erratum Acustica 50, 155–159 (1982)

15.153 P.D. Koopman, C.D. Hanzelka, J.P. Cottingham: Frequency and amplitude of vibration of reeds from American reed organs as a function of pressure, J. Acoust. Soc. Am. **99**, 2506 (1996)

15.154 J.-P. Dalmont, J. Gilbert, J. Kergomard: Reed instruments, from small to large amplitude periodic oscillations and the Helmholtz motion analogy, Acustica **86**, 671–684 (2000)

15.155 J. Backus: Vibrations of the reed and air column of the clarinet, J. Acoust. Soc. Am. **33**, 806–809 (1961)

15.156 N.H. Fletcher: Nonlinear theory of musical wind instruments, Appl. Acoust. **30**, 85–115 (1990)

15.157 R.T. Schumacher: Self-sustained oscillations of a clarinet; an integral equation approach, Acustica **40**, 298–309 (1978)

15.158 R.T. Schumacher: Ab initio calculations of the oscillations of a clarinet, Acustica **48**, 71–85 (1981)

15.159 J. Gilbert, J. Kergomard, E. Ngoya: Calculation of steady state oscillations of a clarinet using the harmonic balance technique, J. Acoust. Soc. Am. **86**, 35–41 (1989)

15.160 D.M. Campbell: Nonlinear dynamics of musical reed and brass wind instruments, Contemp. Phys. **40**, 415–431 (1999)

15.161 N. Gand, J. Gilbert, F. Lalöe: Oscillation threshold of wood-wind instruments, Acta Acustica **1**, 137–151 (1997)

15.162 J.-P. Dalmont, J. Kergomard: Elementary model and experiments for the Helmholtz motion of single reed wind instruments. In: *Proc. Int. Symp. Mus. Acoust.* (IRCAM, Paris 1995) pp. 115–120

15.163 A.H. Benade: Air column reed and player's windway interaction in musical instruments. In: *Vocal Fold Physiology, Biomechanics, Acoustics and Phonatory Control*, ed. by I.R. Titze, R.C. Scherer (Denver Centre for the Performing Arts, Denver 1985) pp. 425–452

15.164 G.P. Scavone: *Modelling vocal tract influence in reed wind instruments*, *Proc. SMAC 03* (Royal Swedish Academy of Music, Stockholm 2003) pp. 291–294

15.165 N.H. Fletcher: Mode locking in non-linearly excited inharmonic musical oscillators, J. Acoust. Soc. Am. **64**, 1566–1569 (1978)

15.166 D. Ayers: *Basic tests for models of the lip reed*, *ISMA 2001* (Fondazione Scuola Di San Giorgio, Venice 2001) pp. 83–86

15.167 S.J. Elliot, J.M. Bowsher: Regeneration in brass wind instruments, J. Sound Vibrat. **83**, 181–207 (1982)

15.168 F.C. Chen, G. Weinreich: Nature of the lip-reed, J. Acoust. Soc. Am. **99**, 1227–1233 (1996)

15.169 S. Adachi, M. Sato: Trumpet sound simulation using a two-dimensional lip vibration model, J. Acoust. Soc. Am. **99**, 1200–1209 (1996)

15.170 I.R. Titze: The physics of small-amplitude oscillation of the vocal folds, J. Acoust. Soc. Am. **83**, 1536–1552 (1988)

15.171 D.C. Copley, W.J. Strong: A stroboscopic study of lip vibrations in a trombone, J. Acoust. Soc. Am. **99**, 1219–1226 (1996)

15.172 S. Yoshikawa, Y. Muto: *Brass player's skill and the associated li-p wave propagation*, ISMA 2001 (Fondazione Scuola Di San Giorgio, Venice 2002) pp. 91–949

Amazon EX1014 Page 674

15.173 J. Gilbert, S. Ponthus, J.F. Petoit: Artificial buzzing lips and brass instruments: experimental results, J. Acoust. Soc. Am. **104**, 1627–1632m (1998)

15.174 J. Cullen, J.A. Gilbert, D.M. Campbell: Brass instruments: linear stability analysis and experiments with an artificial mouth, Acustica **86**, 704–724 (2000)

15.175 N.H. Fletcher: Nonlinear theory of musical wind instruments, Appl. Acoust. **30**, 85–115 (1990)

15.176 T.H. Long: The performance of cup-mouthpiece instruments, J. Acoust. Soc. Am. **19**, 892–901 (1947)

15.177 A. Hirschberg, J. Gilbert, R. Msallam, A.P.J. Wijnands: Shock waves in trombones, J. Acoust. Soc. Am. **99**, 1754–1758 (1996)

15.178 M.E. McIntyre, R.T. Schumacher, J. Woodhouse: On the oscillations of musical instruments, J. Acoust. Soc. Am. **74**, 1325–1345 (1983)

15.179 R.D. Ayers: Impulse responses for feedback to the driver of a musical wind instrument, J. Acoust. Soc. Am. **100**, 1190–1198 (1996)

15.180 B. Fabre, A. Hirschberg: Physical modelling of flue instruments: a review of lumped models, Acustica **86**, 599–610 (2000)

15.181 N.H. Fletcher: Sound production in organ pipes, J. Acoust. Soc. Am. **60**, 926–936 (1976)

15.182 P. Savic: On acoustically effective vortex motions in gaseous jets, Philos. Mag. **32**, 287–252 (1941)

15.183 J.W. Coltman: Jet drive mechanism in edge tones and organ pipes, J. Acoust. Soc. Am. **60**, 723–733 (1976)

15.184 S. Adachi: *CFD analysis of air jet deflection-comparison with Nolle's measurements, SMAC 03* (Royal Swedish Academy of Music, Stockholm 2003) pp. 313–319

15.185 A.W. Nolle, Sinuous instability of a planar air jet: Propagation parameters and acoustic excitation, J. Acoust. Soc. Am. **103**, 3690–3705 (1998)

15.186 L. Cremer, H. Ising: Die selbsterregten Schwingungen von Orgelpfeifen, Acustica **19**, 143–153 (1967)

15.187 J.W. Coltman: Sounding mechanism of the flute and organ pipe, J. Acoust. Soc. Am. **44**, 983–992 (1968)

15.188 M. Raffel, C. Willert, J. Kompenhans: *Particle Image Velocimetry – A Practical Guide* (Springer, Berlin, Heidelberg 1998)

15.189 C. Ségoufin, B. Fabre, M.P. Verge, A. Hirschberg, A.P.J. Wijnands: Experimental study of the influence of mouth geometry on sound production in recorder-like instruments: windway length and chamfers, Acta Acustica **86**, 649–661 (2000)

15.190 S. Thwaites, N.H. Fletcher: Wave propagation on turbulent jets, Acustica **45**, 175–179 (1980)

15.191 S. Thwaites, N.H. Fletcher: Wave propagation of turbulent jets: II, Acustica **51**, 44–49 (1982)

15.192 B. Fabre, A. Hirschberg: *From sound synthesis to instrument making: an overview of recent researches on woodwinds, SMAC 03* (Royal Swedish Academy of Music, Stockholm 2003) pp. 239–242

15.193 N.H. Fletcher: Jet-drive mechanism in organ pipes, J. Acoust. Soc. Am. **60**, 481–483 (1976)

15.194 S.A. Elder: On the mechanism of sound production in organ pipes, J. Acoust. Soc. Am. **54**, 1554–1564 (1973)

15.195 R.C. Chanaud: Aerodynamic whistles, Sci. Am. **222**(1), 40–46 (1970)

15.196 T.A. Wilson, G.S. Beavers, M.A. DeCoster, D.K. Holger, M.D. Regenfuss: Experiments on the fluid mechanics of whistling, J. Acoust. Soc. Am. **50**, 366–372 (1971)

15.197 D.K. Holger, T. Wilson, G. Beavers: Fluid mechanics of the edge-tone, J. Acoust. Soc. Am. **62**, 1116–1128 (1977)

15.198 S. Yoshikawa: Jet wave amplification in organ pipes, J. Acoust. Soc. Am. **103**, 2706–2717 (1998)

15.199 B. Fabre, A. Hirschberg, A.P.J. Wijnands: Vortex shedding in steady oscillations of a flue organ pipe, Acustica – Acta Acustica **82**, 877–883 (1996)

15.200 A. Baines: *European and American Musical Instruments* (Chancellor, London 1983)

15.201 T.D. Rossing: *Science of Percussion Instruments* (World Scientific, Singapore 2000)

15.202 J. Blades: *Percussion Instruments and Their History* (Faber, Faber, London 1974)

15.203 P. Boulez: *Notations I–IV* (Universal Edition, Vienna 1945, 78, 84)

15.204 M. Bertsch: *Vibration patterns and sound analysis of the Viennese timpani. ISMA 2001* (Fondazione Scuola Di San Giorgio, Venice 2001)

15.205 C.V. Raman: The Indian Musical Drum, Proc. Indian Acad. Sci. A1 179 (1934). In: *Reprinted in Musical Acoustics Selected Reprints*, ed. by T.D. Rossing (Am. Assoc. Phys. Teach., College Park 1988)

15.206 T.D. Rossing, W.A. Sykes: Acoustics of indian drums, Percuss. Notes **19**(3), 58 (1982)

15.207 B.S. Ramakrishna, M.M. Sondhi: Vibrations of indian musical drums regarded as composite membranes, J. Acoust. Soc. Am. **26**, 523 (1954)

15.208 S. De: Experimental study of the vibrational characteristics of a loaded kettledrum, Acustica **40**, 206 (1978)

15.209 H. Zhao: *Acoustics of Snare Drums: An Experimental Study of the Modes of Vibration, Mode Coupling and Sound Radiation Pattern, M.S. thesis* (Northern Illinois Univ., DeKalb 1990)

15.210 C.V. Raman: On some Indian stringed instruments, Proc. Indian Assoc. Adv. Sci. **7**, 29–33 (1922)

15.211 T.D. Rossing, I. Bork, H. Zhao, D. Fystrom: Acoustics of snare drums, J. Acoust. Soc. Am. **92**, 84–94 (1992)

15.212 T.D. Rossing: Acoustics of percussion instruments: Part I, Phys. Teacher **14**, 546–556 (1976)

15.213 T.D. Rossing: Chladni's law for vibrating plates, Am. J. Phys. **50**, 271–274 (1982)

Amazon EX1014 Page 675

15.214 N.H. Fletcher: Nonlinear frequency shifts in quasi-spherical-cap shells: Pitch glide in Chinese Gongs, J. Acoust. Soc. Am. **78**, 2069 (1985)
15.215 P.L. Grossman, B. Koplik, Y.-Y. Yu: Nonlinear vibration of shallow spherical shells, J. Appl. Mech. **36**, 451–458 (1969)
15.216 K.A. Legge, N.H. Fletcher: Non-linear mode coupling in symmetrically kinked bars, J. Sound Vib. **118**, 23–34 (1987)
15.217 S. Schedin, P.O. Gren, T.D. Rossing: Transient wave response of a cymbal using double pulsed TV holography, J. Acoust. Soc. Am. **103**, 1217–1220 (1998)
15.218 R. Perrin, T. Charnley, H. Banu, T.D. Rossing: Chladni's law and the modern English church bell, J. Sound Vib. **102**, 11–19 (1985)
15.219 R. Perrin, T. Charnley, J. de Pont: Normal modes of the modern English church bell, J. Sound Vib. **102**, 28048 (1983)
15.220 T.D. Rossing, R. Perrin: Vibration of bells, Appl. Acoust. **20**, 41–70 (1988)
15.221 R. Perrin, T. Charnley, J. de Pont: Normal modes of the modern English church bell, J. Sound Vibr. **102**, 28048 (1983)
15.222 T.D. Rossing, D.S. Hampton, B.E. Richardson, H.J. Sathoff: Vibrational modes of Chinese two-tone bells, J. Acoust. Soc. Am. **83**, 369–373 (1988)

Amazon EX1014 Page 676

16. The Human Voice in Speech and Singing

This chapter describes various aspects of the human voice as a means of communication in speech and singing. From the point of view of function, vocal sounds can be regarded as the end result of a three stage process: (1) the compression of air in the respiratory system, which produces an exhalatory airstream, (2) the vibrating vocal folds' transformation of this air stream to an intermittent or pulsating air stream, which is a complex tone, referred to as the voice source, and (3) the filtering of this complex tone in the vocal tract resonator. The main function of the respiratory system is to generate an overpressure of air under the glottis, or a subglottal pressure. Section 16.1 describes different aspects of the respiratory system of significance to speech and singing, including lung volume ranges, subglottal pressures, and how this pressure is affected by the ever-varying recoil forces. The complex tone generated when the air stream from the lungs passes the vibrating vocal folds can be varied in at least three dimensions: fundamental frequency, amplitude and spectrum. Section 16.2 describes how these properties of the voice source are affected by the subglottal pressure, the length and stiffness of the vocal folds and how firmly the vocal folds are adducted. Section 16.3 gives an account of the vocal tract filter, how its form determines the frequencies of its resonances, and Sect. 16.4 gives an account for how these resonance frequencies or formants shape the vocal sounds by imposing spectrum peaks separated by spectrum valleys, and how the frequencies of these peaks determine vowel and voice qualities. The remaining sections of the chapter describe various aspects of the acoustic signals used for vocal communication in speech and singing. The syllable structure is discussed in Sect. 16.5, the closely related aspects of rhythmicity and timing in speech and singing is described in Sect. 16.6, and pitch and rhythm aspects in Sect. 16.7. The impressive control of all these acoustic characteristics of vocal signals is discussed in Sect. 16.8, while Sect. 16.9 considers expressive aspects of vocal communication.

16.1 **Breathing** 669
16.2 **The Glottal Sound Source** 676
16.3 **The Vocal Tract Filter** 682
16.4 **Articulatory Processes, Vowels and Consonants** 687
16.5 **The Syllable** 695
16.6 **Rhythm and Timing** 699
16.7 **Prosody and Speech Dynamics** 701
16.8 **Control of Sound in Speech and Singing** 703
16.9 **The Expressive Power of the Human Voice** 706
References 706

16.1 Breathing

The process of breathing depends both on mechanical and muscular forces (Fig. 16.1).

During *inspiration* the volume of the chest cavity is expanded and air rushes into the lungs. This happens mainly because of the contraction of the *external intercostals* and the *diaphragm*. The external intercostal muscles raise the ribs. The diaphragm which is the dome-shaped muscle located below the lungs, flattens on contraction and thus lowers the floor of the thoracic cavity.

The respiratory structures form an elastic mechanical system that produces expiratory or inspiratory subglottal pressures, depending on the size of the lung volume, Fig. 16.2. Thus, exhalation and inhalation will always produce forces whose effect is to move the ribs and the lungs back to their resting state, often referred

Amazon EX1014 Page 677

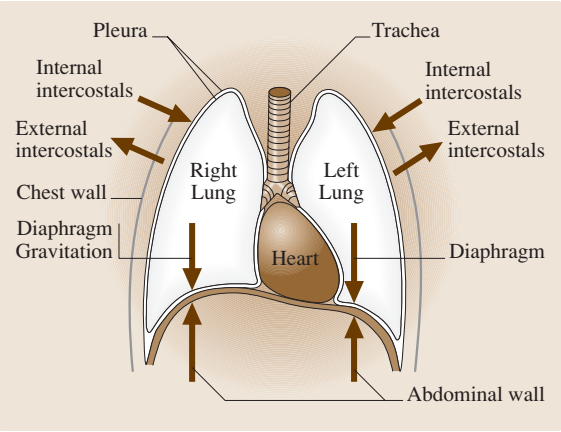


Fig. 16.1 Directions of the muscular and elastic forces. The direction of the gravitation force is applicable to the upright position

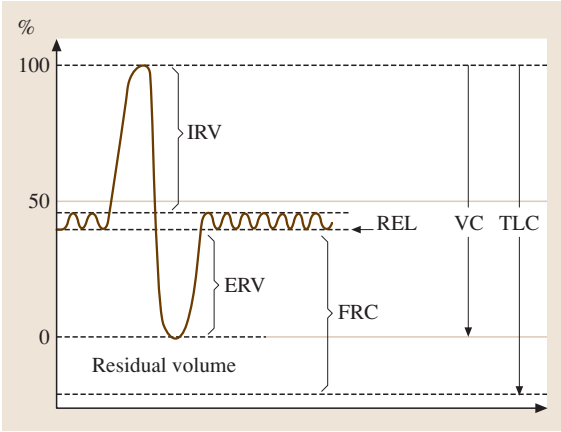


Fig. 16.3 Definition of the various terms for lung volumes. The graph illustrates the lung volume changes during quiet breathing interrupted by a maximal inhalation followed by a maximal exhalation. VC is the vital capacity, TLC is the total lung capacity, IRV and ERV are the inspiratory and expiratory reserve volume, REL is the resting expiratory level, FRC is the functional residual capacity

to as the resting expiratory level (REL). The deeper the breath, the greater this force of *elastic recoil*. This component plays a significant part in pushing air out of the lungs, both in speech and singing, and especially at large lung volumes. The elasticity forces originate both from the rib cage and the lungs. As illustrated in Fig. 16.2, the rib cage produces an exhalatory force at high lung volumes and an inhalatory force at low lung volumes, and the lungs always exert an exhalatory force. As a consequence, activation of inspiratory muscles is needed for producing a low subglottal pressure, e.g. for singing a soft (pianissimo) tone, at high lung volume. Conversely, activation of expiratory muscles is needed for producing a high subglottal pressure, e.g. for singing a loud (fortissimo) tone, at low lung volume.

In addition to mechanical factors, exhaling may involve the activity of the *internal intercostals* and the *abdominal muscles*. Contraction of the former has the effect of lowering the ribs and thus compressing the chest cavity. Activating the abdominal muscles generates upward forces that also contribute towards reducing the volume of the rib cage and the lungs. The function of these muscles is thus expiratory.

Another significant factor is gravity whose role depends on body posture. In an upright position, the diaphragm and adjacent structures tend to be pulled down, increasing the volume of the thoracic cavity. In this sit-

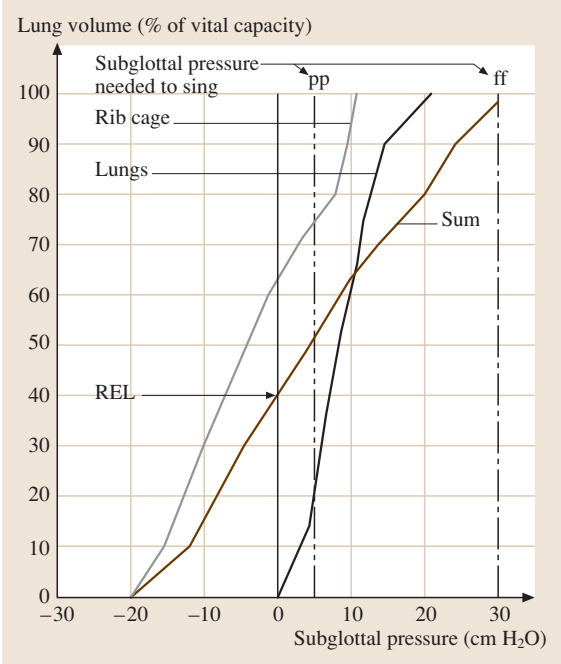


Fig. 16.2 Subglottal pressures produced at different lung volumes in a subject by the recoil forces of rib cage and lungs. The resting expiratory level (REL) is the lung volume at which the inhalatory and the exhalatory recoil forces are equal. The *thin* and *heavy chain-dashed lines* represent subglottal pressures typically needed for soft and very loud phonation

Amazon EX1014 Page 678

uation, the effect of gravity is inspiratory. In contrast, in the supine position the diaphragm tends to get pushed up into the rib cage, and expiration is promoted [16.1].

The total air volume that is contained in a maximally expanded rib cage is called the *total lung capacity* (TLC in Fig. 16.3). After maximum exhalation a small air volume, the *residual volume*, is still left in the airways. The greatest air volume that can be exhaled after a maximum inhalation is called the *vital capacity* (VC) and thus equals the difference between the TLC and the residual volume. The lung volume at which the exhalatory and inhalatory recoil forces are equal, or REL, is reached after a relaxed sigh, see Figs. 16.2 and 16.3. During tidal breathing inhalation is initiated from REL, so that inhalation is active resulting from an activation of inspiratory muscles, and exhalation is passive, produced by the recoil forces. In tidal breathing only some 10% of VC is inhaled, such that a great portion of the VC, the *inspiratory reserve volume*, is left. The air volume between REL and the residual volume is called the *expiratory reserve volume*.

VC varies depending on age, body height and gender. At the age of about 20 years, an adult female has a vital capacity of 3–3.6 l depending on body height, and for males the corresponding values are about 4–5.5 l.

Experimental data [16.4] show that, during tidal breathing, lung volume variations are characterized by a regular quasi-sinusoidal pattern with alternating segments of inspiration and expiration of roughly equal duration (Fig. 16.4). In speech and singing, the pattern is transformed. Inspirations become more rapid and expiration occurs at a slow and relatively steady rate. Increasing vocal loudness raises the amplitude of the lung volume records but leaves its shape relatively unchanged.

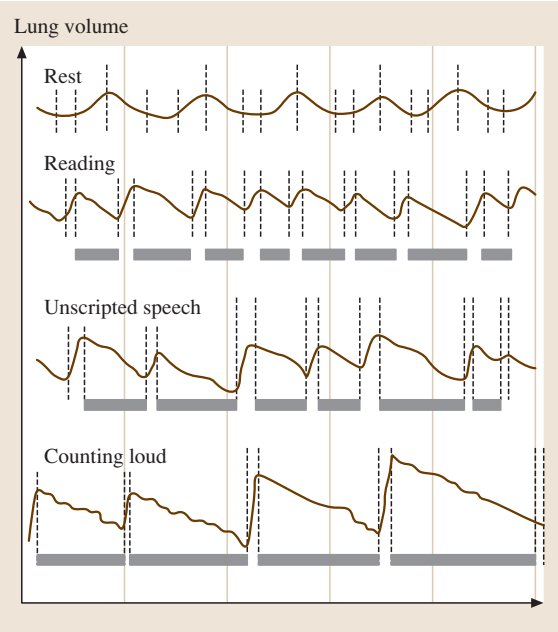


Fig. 16.4 Examples of speech breathing

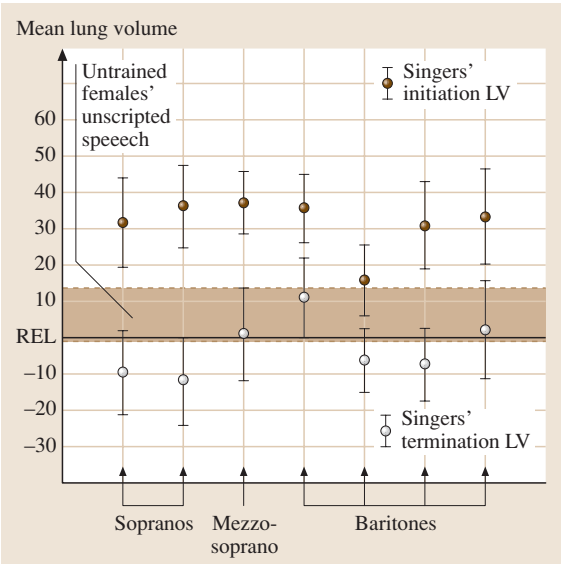


Fig. 16.5 Lung volume averages used in speech and operatic singing expressed as percentages of vital capacity relative to the resting expiratory level REL. The shaded band represents the mean lung volume range observed in untrained female voices' unscripted speech (after [16.2]). The filled and open symbols show the corresponding measures for professional opera singers of the indicated classifications when performing opera arias according to *Thomasson* [16.3]. The bars represent ± one SD

Figure 16.5 shows mean lung volumes used by professional singers when singing well-rehearsed songs. The darker band represents the mean lung volume range observed in spontaneous speech [16.2]. The lung volumes of conversational speech are similar to those in tidal breathing. Loud speech shows greater air consumption and thus higher volumes (Fig. 16.4). Breath groups in speech typically last for 3–5 seconds and are terminated when lung volumes approach the relaxation expiratory level REL, as illustrated in Fig. 16.5. Thus, in phonation, lung volumes below REL are mostly avoided.

Amazon EX1014 Page 679

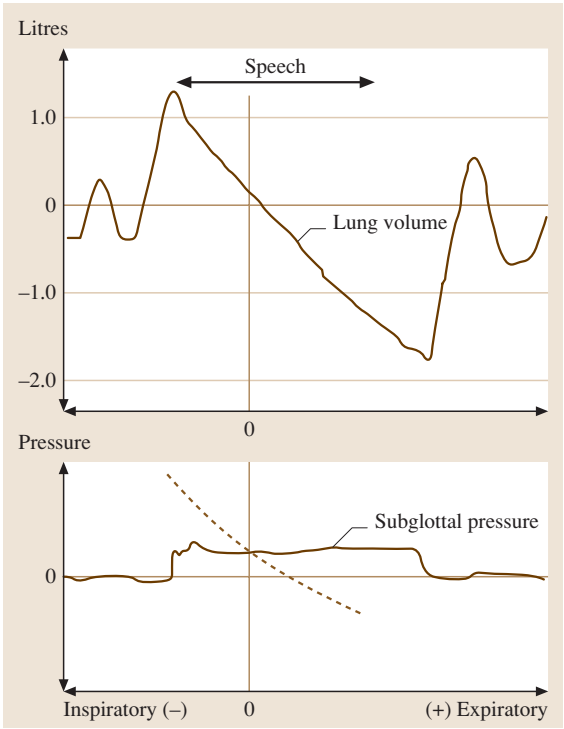


Fig. 16.6 The records show lung volume (relative to the mid-respiratory level), subglottal (esophageal) pressure and stylized muscular activity for speaker counting from one to 32 at a conversational vocal effort. (After Draper et al. [16.5]) . To the *left* of the vertical line recoil forces are strongly expiratory, to the *right* they are inspiratory. *Arrows* have been added to the *x*-axis to summarize the original EMG measurements which indicate that the recoil forces are balanced by muscular activity (EMG = electromyography, measurement of the electrical activity of muscles). To the *left of the vertical line* (as indicating by *left-pointing arrow*) the net muscular force is inspiratory, to the *right* it is expiratory (*right-pointing arrow*). To keep loudness constant the talker maintains a stable subglottal pressure and recruits muscles according to the current value of the lung volume. This behavior exemplifies the phenomenon known as *motor equivalence*

In singing, breath groups tend to be about twice as long or more, and air consumption is typically much greater than in conversational speech. Mostly they are terminated close to the relaxation expiratory level, as in speech, but sometimes extend into the *expiratory reserve volume* as illustrated in Fig. 16.5. This implies that, in singing, breath groups typically start at much higher lung volumes than in speech. This use of high lung volumes implies that singers have to deal with much greater recoil forces than in speech.

Figure 16.6 replots, in stylized form, a diagram published by Ladefoged et al. [16.7]. It summarizes measurements of lung volume and subglottal pressure, henceforth referred to as Ps, recorded from a speaker asked to take a deep breath and then to start counting. The dashed line intersecting the Ps record represents the relaxation pressure. This line tells us that elastic recoil forces are strongly expiratory to the left of the vertical line. To the right they produce a pressure lower than the target pressure and eventually an inspiratory pressure. The Ps curve remains reasonably flat throughout the entire utterance.

The question arises: how can this relative constancy be achieved despite the continually changing contribution of the relaxation forces? In a recent replication of this classical work [16.8], various criticisms of the original study are addressed but basically the original answer to this question is restated: the motor system adapts to the external goal of keeping the Ps fairly constant (Fig. 16.6). Initially, when recoil forces are strong, muscle activity is predominantly in the *inspiratory* muscles such as the *diaphragm* and the *external intercostals*. Gradually, as recoil forces decline, the *ex-*

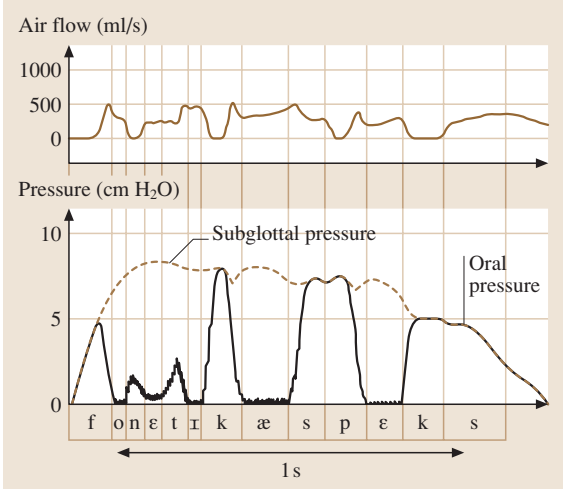


Fig. 16.7 Example of oral and subglottal pressure variations for the phrase "*phonetic aspects*" (after *Netsell* [16.6]). As the glottis opens and closes for voiceless and voiced sounds and the vocal tract opens or closes for vowels and consonants, the expired air is opposed by varying degrees of glottal and supraglottal impedance. Both the oral and the subglottal pressure records reflect the combined effect of these resistance variations

Amazon EX1014 Page 680

piratory muscles (the *internal intercostals*, the *rectus abdominis* among others) take over increasingly as the other group relaxes (cf. arrows, Fig. 16.6). According to our present understanding [16.8, 9], this adaptation of breathing for speech is achieved by constantly updating the balance between agonist and antagonist muscles in accordance with current conditions (lung volume, body posture, etc.) and in order to meet the goal of constant Ps. Muscle recruitment depends on lung volume.

Figure 16.7 presents a representative example of oral and Ps records. The phrase is "*phonetic aspects*" (After *Netsell* [16.6]). The top panel shows a record of oral air flow. Vertical lines indicate acoustic segment boundaries. The bottom diagram superimposes the curves for oral and subglottal pressure.

The Ps shows a falling pattern which becomes more pronounced towards the end of the phrase and which is reminiscent of the declination pattern of the fundamental frequency contour typical of declarative sentences [16.10, p. 127]. The highest values occur at the midpoints of [ɛ] and [æ], the stressed vowels of the utterance. For vowels, oral pressure is close to atmospheric pressure (near zero on the *y*-axis cm H_2O scale). The [k] of *phonetic* and the [p] of *aspects* show highly similar traces. As the tongue makes the closure for [k], the air flow is blocked and the trace is brought down to zero. This is paralleled by the oral pressure rising until it equals the Ps. As the [k] closure progresses, a slight increase builds up in both curves. The release of the [k] is signaled by a peak in the air flow and a rapid decrease in Ps. An almost identical pattern is seen for [p].

In analyzing Ps records, phoneticians aim at identifying variations based on an *active control* of the respiratory system and phenomena that can be attributed to the system's *passive response* to ongoing activity elsewhere, e.g., in the vocal tract and or at the level of the vocal folds [16.11].

To exemplify passive effects let us consider the events associated with [k] and [p] just discussed. As suggested by the data of Figs 16.4 and 16.6, speech breathing proceeds at a relatively steady lung volume decrement. However, the open or closed state of the glottis, or the presence of a vocal tract constriction/closure, is capable of creating varying degrees of impedance to the expired air. The oral pressure record reflects the combined effect of glottal and articulatory resistance variations. Ps is also affected by such changing conditions. As is evident from Fig. 16.7, the Ps traces during the [k] and [p] segments first increase during the stop closures. Then they decrease rapidly during the release and the aspiration phases. These effects are passive responses to the segment-based changes in supraglottal resistance and are unlikely to be actively programmed [16.12, 13].

Does respiration play an active role in the production of stressed syllables? In "*phonetic aspects*" main stress occurs on [ɛ] and [æ]. In terms of Ps, these vowels exhibit the highest values. Are they due to an active participation of the respiratory system in signaling stress, or are they fortuitous by-products of other factors?

An early contribution to research on breathing and speech is the work by *Stetson* [16.14]. On the basis of aerodynamic, electromyographic and chest movement measurements, Stetson proposed the notion of the *chest pulse*, a chunk of expiratory activity corresponding to the production of an individual syllable.

In the late fifties, Ladefoged and colleagues published an electromyographic study [16.7] which cast doubt on Stetson's interpretations. It reported increased activity in expiratory muscles (*internal intercostals*) for English syllables. However, it failed to support Stetson's chest pulse idea in that the increases were found only on stressed syllables. *Ladefoged* [16.8] reports findings from a replication of the 1958 study in which greater activity in the internal intercostals for stressed syllables was confirmed. Moreover, reduced activity in inspiratory muscles (*external intercostals*) was observed to occur immediately before each stressed syllable.

Ps measurements provide further evidence for a positive role for respiration in the implementation of stress. *Ladefoged* [16.15, p. 143] states: "Accompanying every stressed syllable there is always an increase in the Ps". This claim is based on data on disyllabic English noun–verb pairs differing in the position of the main stress: *TORment* (noun)–*torMENT* (verb), *INsult* (noun)–*inSULT* (verb) etc. A clear distinction between the noun and verb forms was observed, the former having Ps peaks in the first syllable and the latter on the second syllable.

As for segment-related Ps variations (e.g., the stops in Fig. 16.7), there is wide agreement that such local perturbations are induced as automatic consequences of speech production aerodynamics. Explaining short-term ripple on Ps curves in terms of aerodynamics is consistent with the observation that the respiratory system is mechanically sluggish and therefore less suited to implement rapid Ps changes in individual phonetic segments. Basically, its primary task is to produce a Ps contour stable enough to maintain vocal intensity at a fairly constant level (Fig. 16.6).

Amazon EX1014 Page 681

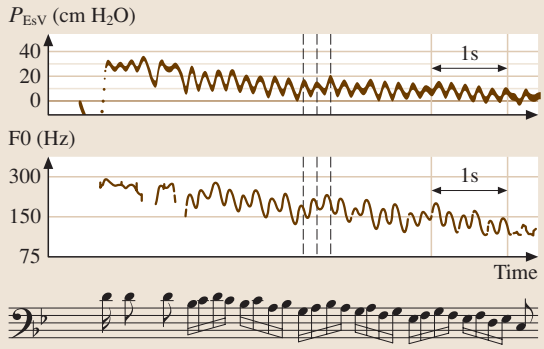


Fig. 16.8 Changes in esophageal pressure, corresponding to subglottic pressure changes (*upper curve*) and phonation frequency (*middle*) in a professional baritone singer performing the coloratura passage shown at the bottom. (After *Leanderson* et al. [16.17])

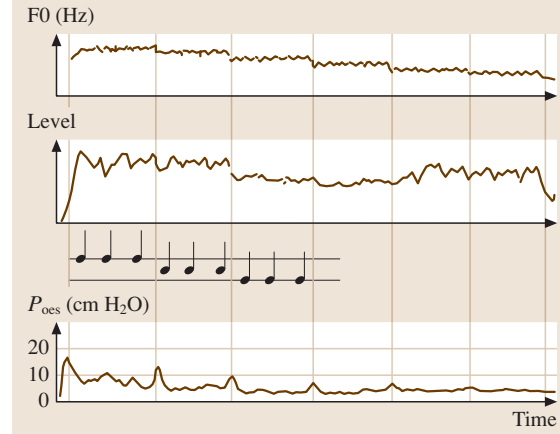


Fig. 16.9 The three *curves* show, from *top to bottom* F0 sound level and esophageal pressure in a professional baritone singer singing the music example shown in the *graph*, i. e. a descending scale with three tones on each scale tone. The singer marked the first beat in each bar by a pressure pulse. (After *Sundberg* et al. [16.18])

Singing provides experimental data on the behavior of the respiratory system that tend to reinforce the view that a time varying respiration activity can be an active participant in the sound generation process. Although there is ample justification for saying that the mechanical response of the respiratory structures is characterized by a long time constant, Ps in singing varies quickly and accurately. One example is presented in Fig. 16.8. It shows variations in pressure, captured as the pressure variations in the esophagus and thus mirroring the Ps variations [16.16], during a professional baritone singer's performance of the music example shown at the bottom. Each tone in the sequence of sixteenth notes is produced with a pressure pulse. Note also that in the fundamental frequency curve each note corresponds to a small rise and fall. The tempo of about six sixteenth notes per second implies that the duration of each rise–fall cycle is about 160 ms. It would take quite special skills to produce such carefully synchronized Ps and fundamental frequency patterns.

Synthesis experiments have demonstrated that this particular fundamental frequency pattern is what produces what is perceived as a sung legato coloratura [16.19, 20]. The voice organ seems incapable of producing a stepwise-changing F0 curve in legato, and such a performance therefore sounds as if it was produced by a music instrument rather than by a voice. In this sense this F0 variation pattern seems to be needed for eliciting the perception of a sung sequence of short legato tones.

The pressure variations are not likely to result from a modulation of glottal adduction. A weak glottal adduction should result in a voice source dominated by the fundamental and with very weak high spectrum partials, and mostly the amplitudes of the high overtones do not vary in coloratura sequences.

It is quite possible that the F0 variations are caused by the Ps variations. In the example shown in Fig. 16.8, the pressure variations amount to about 10 cm H_2O, which should cause a F0 modulation of about 30 Hz or so, corresponding to two semitones in the vicinity of 220 Hz. This means that the coloratura pattern may simply be the result of a ramp produced by the F0 regulating system which is modulated by the pressure pulses produced by the respiratory system.

As another example of the skilled use of Ps in singing, Fig. 16.9 shows the pressure variation in the esophagus in a professional baritone singer performing a descending scale pattern in 3/4 time, with three tones on each scale step as shown by the score fragment in the first three bars. The singer was instructed to mark the first tone in each bar. The pressure record demonstrates quite clearly that the first beat was produced with a marked pulse approximately doubling the pressure. Pulses of this magnitude must be produced by the respiratory apparatus. When the subject was instructed to avoid marking the first tone in each bar, no pressure pulses were observed [16.18].

The effect of Ps on fundamental frequency has been investigated in numerous studies. A commonly used

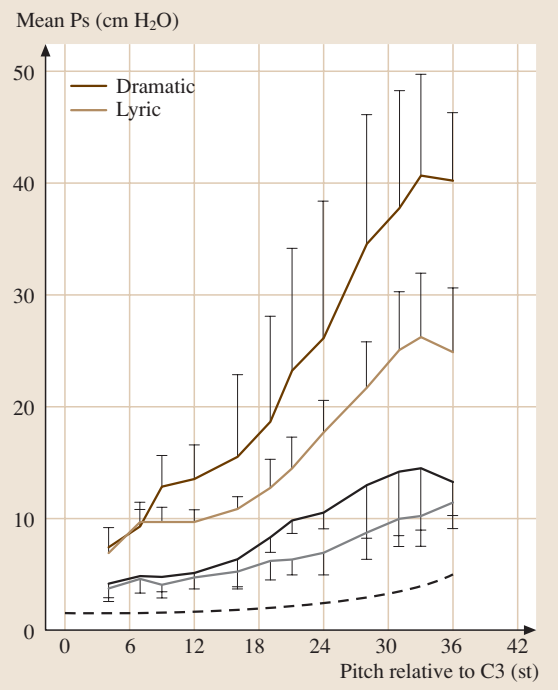


Fig. 16.10 Mean subglottal pressure, captured as the oral pressure during /p/-occlusion, in five lyric and six dramatic professional sopranos (*heavy* and *thin curves*, respectively) who sang as loudly and as softly as possible at different F0. The *dashed curve* shows Titze's prediction of threshold pressure, i. e., the lowest pressure that produces vocal fold vibration. The *bars* represent one SD (After [16.21])

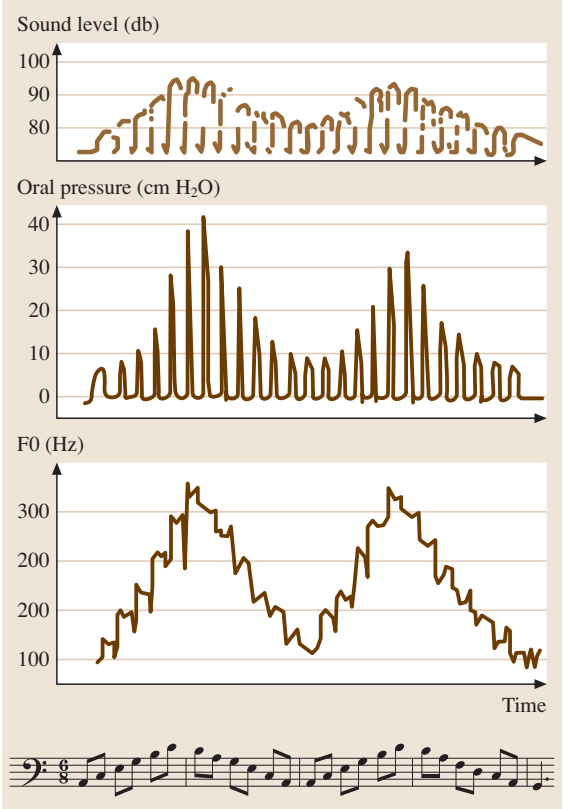


Fig. 16.11 Simultaneous recordings of sound level, subglottic pressure (captured as the oral pressure during /p/ occlusion) and F0 in a professional singer singing the note example shown at the bottom. (After *Leanderson* et al. [16.17])

method is to ask the subject to produce a sustained vowel at a given steady pitch and then, at unpredictable moments, change the Ps by applying a push to the subject's abdomen or chest. The assumption underlying these studies is that there will be an initial interval during which possible reflex action of laryngeal muscles will not yet come into play. Hence the data from this time segment should give a valid picture of the relationship between fundamental frequency and Ps.

In a study using the push method, *Baer* [16.22] established this relationship for a single male subject. Data on steady phonations at 94, 110, 220 Hz and a falsetto condition (240 Hz) were collected. From electromyograms from the *vocalis* and the *interarytenoid* it was possible to tell that the first 30 ms after the push onset were uncontaminated by reflex muscle responses. Fundamental frequency was plotted against Ps for the four F0 conditions. These plots showed linear data clusters with slopes of 4, 4, 3 and 9 Hz/cm H_2O respectively. Other studies which assumed a longer reflex response latency (100 ms) report that F0's dependence on Ps occurs between 2 to 7 Hz/cm H_2O.

From such observations phoneticians have concluded that, in producing the F0 variations of natural speech, Ps plays only a secondary role. F0 control is primarily based on laryngeal muscle activity. Nonetheless, the falling F0 contours of phrases with statement intonation tend to have Ps contours that are also falling. Moreover, in questions with final F0 increases, the Ps records are higher towards the end of the phrase than in the corresponding statements [16.15].

There is clear evidence that the Ps needs to be carefully adapted to the target fundamental frequency in singing, especially in the high range ([16.23], [16.24,

Amazon EX1014 Page 683

p. 36]). An example is presented in Fig. 16.10 which shows mean Ps for five dramatic and five lyric sopranos when they sang pp and ff tones throughout their range [16.21]. The bars represent one standard deviation. The dashed curve is the threshold pressure, i. e. the lowest pressure that produced vocal fold vibration, for a female voice according to *Titze* [16.25, 26].

The singers mostly used more than twice as high Ps values when they were singing their loudest as compared with their softest tones. The subjects reach up to 40 cm H_2O. This can be compared with values typical of normal adult conversational speech, which tend to occur in the range of 5–10 cm H_2O. However, loud speech and particularly stage speech may occasionally approach the higher values for singing.

Also interesting is the fact that lyric sopranos use significantly lower Ps values than dramatic sopranos both in soft and loud phonation. It seems likely that this difference reflects difference in the mechanical properties of the vocal folds.

Figure 16.11 presents another example of the close link between Ps and fundamental frequency. From top to bottom it plots the time functions for sound level (SPL in dB), pressure (cm H_2O), fundamental frequency (Hz) for a professional baritone singing an ascending triad followed by a descending dominant-seventh triad (according to the score at the bottom). The pressure record was obtained by recording the oral pressure as the subject repeated the syllable [pæ] on each note. During the occlusion of the stop the peak oral pressure becomes equal to the Ps (as discussed in connection with Fig. 16.7). That means that the peak values shown in the middle diagram are good estimates of the actual Ps. It is important to note that, when the trace repeatedly returns to a value near zero, what we see is not the Ps, but the oral pressure reading for the [æ] vowel, which is approximately zero.

The exercise in Fig. 16.11 is often sung staccato, i. e. with short pauses rather than with a /p/ consonant between the tones. During these pauses the singer has to get ready for the next fundamental frequency value and must therefore avoid exhaling the pressurized air being built up in the lungs. Singers do this by opening the glottis between the tones and simultaneously reducing their Ps to zero, so that no air will be exhaled during the silent intervals between the tones. A remarkable fact demonstrated here is that, particularly when sung loudly – so that high Ps values are used – this exercise requires nothing less than a virtuoso mastering of both the timing and tuning of the breathing apparatus and the pitch control process. A failure to reach a target pressure is likely to result in a failure to reach the target F0.

16.2 The Glottal Sound Source

In speech and singing the general method to generate sound is to make a constriction and to let a strong flow of air pass through it. The respiratory component serves as the power source providing the energy necessary for sound production. At the glottis the steady flow of air generated by the respiratory component is transformed into a quasiperiodic series of glottal pulses. In the vocal tract, the glottally modified breath stream undergoes further modifications by the resonance characteristics of the oral, pharyngeal and nasal cavities.

Constrictions are formed at the glottis - by adjusting the separation of the vocal folds – and above the glottis – by positioning the articulators of the vocal tract. As the folds are brought close together, they respond to the air rushing through by initiating an open–close vibration and thereby imposing a quasiperiodic modulation of airflow. Thus, in a manner of speaking, the glottal structures operate as a device that imposes an AC modulation on a DC flow. This is basically the way that *voicing*, the sound source of voiced vowels and consonants and the carrier of intonation and melody, gets made.

A second mechanism is found in the production of *noise*, the acoustic raw materials for voiceless sounds (e.g., [f], [s], [p], [k]). The term refers to irregular turbulent fluctuations in airflow which arise when air comes out from a constriction at a high speed. This process can occur at the glottis – e.g., in [h] sounds, whisper and breathy voice qualities – or at various places of articulation in the vocal tract.

The framework for describing both singing and speech is that of the *source-filter theory of speech production* [16.27, 28]. The goal of this section is to put speech and singing side by side within that framework and to describe how the speaker and the singer coordinate respiration, phonation and articulation to shape the final product: the acoustic wave to be perceived by the listener.

Figure 16.12 [16.29] is an attempt to capture a few key aspects of vocal fold vibrations. At the center a sin-

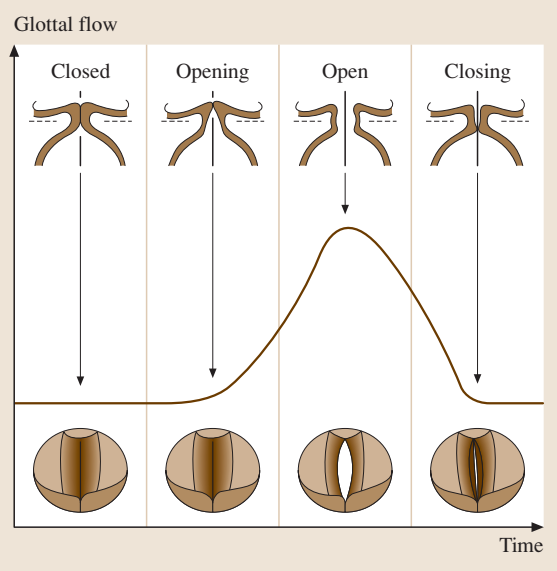


Fig. 16.12 Relationship between the flow glottogram, showing transglottal airflow versus time, and glottal configurations in a coronal plane (*upper series* of images) and from above, as through a laryngeal mirror (*lower series* of images). The airflow increases when the folds open the glottis and allow air to pass and decreases quickly when they close the glottis and arrest the airflow. (After [16.29])

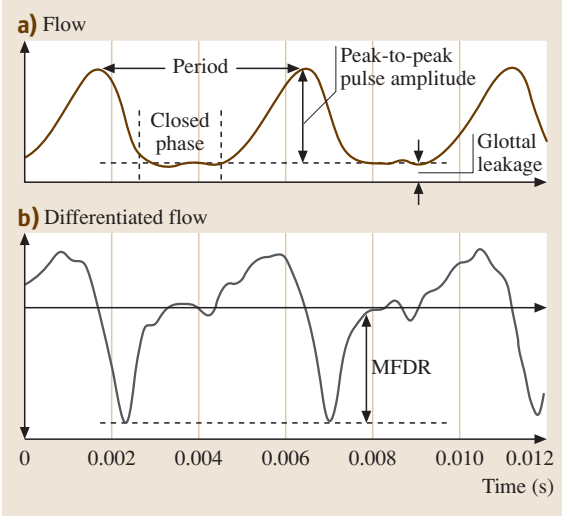


Fig. 16.13a,b Illustration of measures commonly used to characterize a flow glottogram **(a)** and its time derivative, the differentiated flow glottogram **(b)**. (The *wiggles* in the latter are artifacts caused by an imperfect inverse filtering)

gle cycle of a glottal waveform is seen. It plots airflow through the glottis as a function of time. Alternatively, the graph can be used to picture the time variations of the glottal area which present a pattern very similar to that for airflow. The top row shows stylized cross sections of the vocal folds at selected time points during the glottal cycle. From left to right they refer to the opening of the folds, the point of maximum area and the point of closure. Below is a view of the vocal folds from above corresponding to the profiles at the top of the diagram.

There are a number of different methods for visualizing vocal fold vibrations. By placing an electrode on each side of the thyroid cartilage, a minute current can be transferred across the glottis. This current increases substantially when the folds make contact. The resulting *electroglottogram*, also called a *laryngogram*, thus shows how the contact area varies with time. It is quite efficient in measurement of F0 and closed phase. *Optical glottograms* are obtained by illuminating the trachea from outside by means of a strong light source and capturing the light traveling through the glottis by means of an optical sensor in the pharynx. The signal therefore reflects the glottal area, but only as long as the light successfully finds it way to the sensor. A posterior tilting of the epiglottis may easily disturb or eliminate the signal.

Flow glottograms show transglottal airflow versus time and are derived by inverse filtering the audio signal, often picked up as a flow signal by means of a pneumotachograph mask [16.30]. Inverse filtering implies that the signal is passed through a filter with a transfer function equalling the inverted transfer function of the vocal tract. Therefore correct inverse filtering requires that the inverted resonance peaks of the inverse filter are tuned to the formant frequencies of the vowel being filtered.

As transglottal airflow is zero when the glottis is closed and nonzero when it is open, the flow glottogram is physiologically relevant. At the same time it is a representation of the sound of the voice source.

A typical example of a flow glottogram is given in the upper graph of Fig. 16.13. The classical parameters derived from flow glottograms are the durations of the period and of the closed phase, pulse peak-to-peak amplitude, and glottal leakage. The lower graph shows the differentiated glottogram. The negative peak amplitude is often referred to as the *maximum flow declination rate* (MFDR). As we shall see it has special status in the process of voice production.

In the study of both speech and singing, the acoustic parameter of main relevance is the time variations in sound pressure produced by the vocal system and received by the listener's ears. Theoretically, this signal

is roughly proportional to the derivative of the output airflow at the lips [16.28, 31] and it is related to the derivative of the glottal waveform via the transfer function of the vocal tract. Formally, the excitation signal for voiced sounds is defined in terms of this differentiated signal. Accordingly, in source-filter theory, it is the derivative of glottal flow that represents the source and is applied to the filter or resonance system of the vocal tract. The amplitude of the vocal tract excitation, generally referred to as the *excitation strength*, is quantified by the maximum velocity of flow decrease during vocal-fold closing movement (the MFDR, Fig. 16.13) which is a determinant of the level of the radiated sound. At the moment of glottal closure a drastic modification of the air flow takes place. This change is what generates voicing for both spoken and sung sounds and produces a sound with energy across a wide range of frequencies.

The Liljencrants–Fant (LF) model [16.32, 33] is an attempt to model glottal waveforms using parameters such as fundamental frequency, excitation strength, dynamic leakage, open quotient and glottal frequency (defined by the time period of glottal opening phase). Other proposals based on waveform parameters have been made by Klatt and Klatt [16.34], Ljungqvist and Fujisaki [16.35], Rosenberg [16.36] and Rothenberg et al. [16.37]. A second line of research starts out from assumptions about vocal fold mechanics and applies aerodynamics to simulate glottal vibrations [16.38, 39]. Insights from such work indicate the importance of parameters such as Ps, the adducted/abducted position of vocal folds and their stiffness [16.28].

During the early days of speech synthesis it became clear that the simplifying assumption of a constant voice source was not sufficient to produce high-quality natural-sounding copy synthesis. Experimental work on the voice source and on speech synthesis has shown that, in the course of an utterance, source parameters undergo a great deal of variation. The determinants of this dynamics are in part prosodic, in part segmental. Figure 16.14 [16.32] presents a plot of the time variations of the *excitation strength* parameter (i. e. MFDR) during the Swedish utterance: *Inte i DETta århundrade* [ɪntɪ ˣdɛtːaoːrhəndradɛ]. The upper-case letters indicate that the greatest prominence was on the first syllable of *detta*. Vertical lines represent acoustic segment boundaries.

Gobl collected flow data using the mask developed by Rothenberg [16.30] and applied inverse filtering to obtain records of glottal flow which, after differentiation, enabled him to make measurements of excitation strength and other LF parameters.

Figure 16.14 makes clear that excitation strength is in no way constant. It varies depending on both prosodic and segmental factors. The effect of the segments is seen near the consonant boundaries. As the vocal tract is constricted, e.g., in [d] and [tː], and as transglottal pressure therefore decreases (cf. the pressure records of Fig. 16.7), excitation strength is reduced. In part these variations also occur to accommodate the voicing and the voicelessness of the consonant [16.28]. This influence of consonants on the voice source has been documented in greater detail by *Ni Chasaide* and *Gobl* [16.40] for German, English, Swedish, French, and Italian. Particularly striking effects were observed in the context of voiceless consonants.

Prosodically, we note in Fig. 16.14 that excitation strength exhibits a peak on the contrastively stressed syllable in *detta* and that the overall pattern of the phrase is similar to the falling declination contour earlier mentioned for declarative statements.

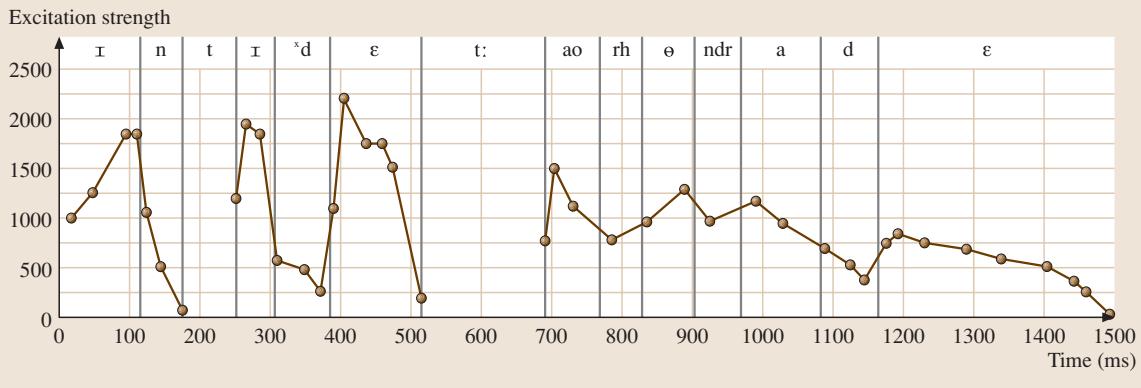


Fig. 16.14 Running speech data on the 'excitation strength parameter' of the LF model. (After [16.32])

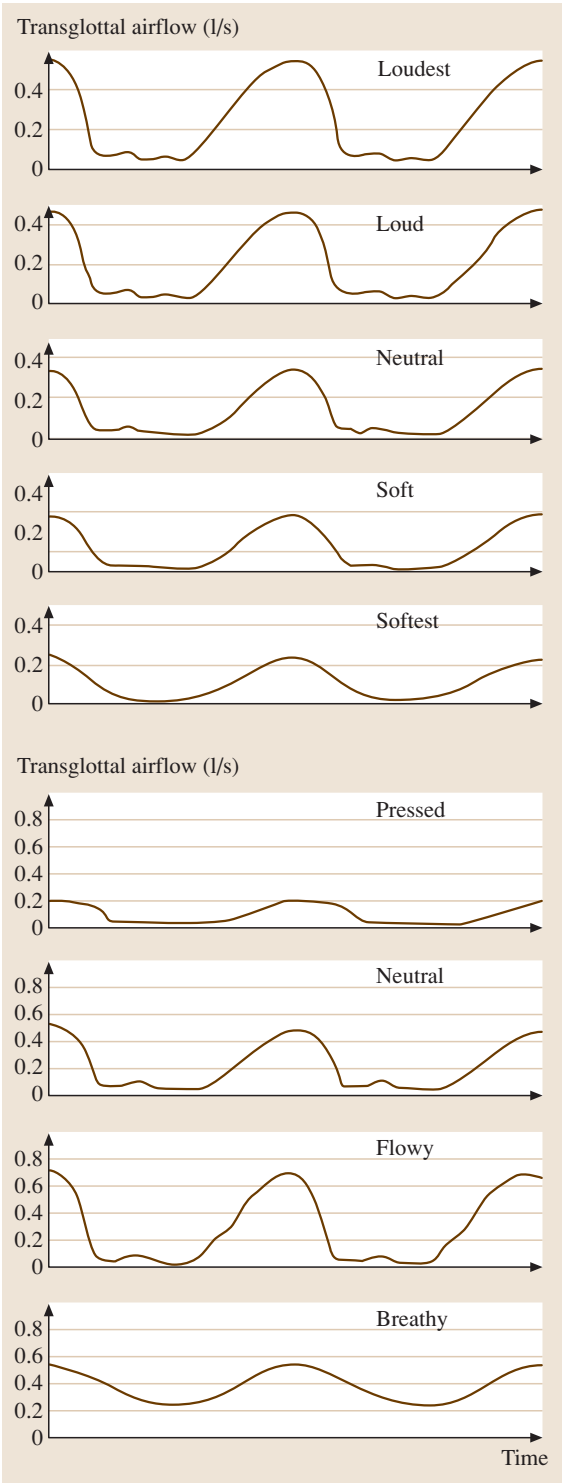


Table 16.1 Measurements of subglottal pressure (cm H_2O) and SPL at 0.3 m (dB) measured for the glottograms shown in Fig. 16.15

Table A	Ps (cm H_2O)	SPL at 0.3 m(dB)
loudest	14.3	84
loud	7.3	82
neutral	5.7	78
soft	3.9	75
softest	3	70
Table B	**Ps (cm H_2O)**	**SPL at 0.3 m(dB)**
pressed	11.4	83
neutral	5.1	79
flowy	8	88
breathy	6.6	84

Examples of the fact that the Ps has a strong influence on the flow glottogram are given in the upper set of graphs of Fig. 16.15, which shows a set of flow glottograms for phonations produced at the same pitch but with varying degrees of *vocal loudness*. As we examine the series of patterns from loudest to softest we note that both the peak flow and the maximum steepness of the trailing end of the pulse, i. e., MFDR, increase significantly with increasing Ps. These shape changes are lawfully related to the variations in Ps and are directly reflected in sound pressure levels as indicated by the numbers in Table 16.1.

Holmberg and colleagues [16.41] made acoustic and airflow recordings of 25 male and 20 female speakers producing repetitions of [pæ] at soft, normal and loud vocal efforts [16.42, p. 136]. Estimates of Ps and glottal airflow were made from recordings of oral pressure and oral airflow. Ps was derived by interpolating between peak oral pressures for successive [p] segments and then averaging across repetitions. A measure of average flow was obtained by low-pass filtering the airflow signal and averaging values sampled at the vowel midpoints.

Fig. 16.15 Typical flow glottogram changes associated with changes of loudness or mode of phonation (*top* and *bottom series*). As loudness of phonation is raised, the closing part of the curve becomes more steep. When phonation is pressed, the glottogram amplitude is low and the closed phase is long. As the adduction force is reduced, pulse amplitude grows and the closed phase becomes shorter (note that the flow scales are different for the *top* and *bottom series* of glottograms. Flowy phonation is the least adducted, and yet not leaky phonation

Amazon EX1014 Page 687

A copy of the airflow signal was low-pass filtered and inverse filtered to remove the effect of F1 and other formants. The output was then differentiated for the purpose of determining the MFDR (Fig. 16.13).

Figure 16.15 also illustrates how the voice source can be continuously varied between different *modes of phonation*. These modes range from hyperfunctional, or pressed, over neutral and flowy to hypofunctional, or breathy. The corresponding physiological control parameter can be postulated to be glottal adduction, i. e. the force by which the folds press against each other. It varies from minimum in hypofunctional to extreme in hyperfunctional. Flowy phonation is produced with the weakest degree of glottal adduction compatible with a full glottal closure. The physiologically relevant property that is affected is the vibration amplitude of the vocal folds, which is small in hyperfunctional/pressed phonation and wide in breathy phonation.

As illustrated in Fig. 16.15 the flow glottogram is strongly affected by these variations in phonation mode [16.43]. In pressed phonation, the pulse amplitude is small and the closed phase is long. It is larger in neutral and even more so in flowy. In breathy phonation, typically showing a waveform similar to a sine wave, airflow is considerable, mainly because of a large leakage, so there is no glottal closure.

Phonation mode affects the relation between Ps and the SPL of the sound produced. As shown in Table 16.1B pressed phonation is less economical from an acoustic point of view: a Ps of 11.4 cm H_2O produces an SPL at 0.3 m of only 83 dB, while in flowy phonation a lower Ps produces a higher SPL.

Pitch, loudness and phonation mode are voice qualities that we can vary continuously. By contrast, vocal registers, also controlled by glottal parameters, appear more like toggles, at least in untrained voices. The voice is operating either in one or another register. There are at least three vocal registers, *vocal fry*, *modal* and *falsetto*. When shifting between the modal and falsetto registers, F0 discontinuities are often observed [16.44].

The definition of vocal registers is quite vague, a set of tones along the F0 continuum that sound similar and are felt to be produced in a similar way. As registers depend on glottal function, they produce different flow glottogram characteristics. Figure 16.16 shows typical examples of flow glottograms for the falsetto and modal registers as produced by professional baritone, tenor and countertenor singers. The pulses are more rounded, the closed phase is shorter, and the glottal leakage is greater in the falsetto than in the modal register. However, the waveform of a given register often varies substantially between individuals. Classically trained sopranos, altos, and tenors learn to make continuous transitions between the modal and the falsetto registers, avoiding abrupt changes in voice timbre.

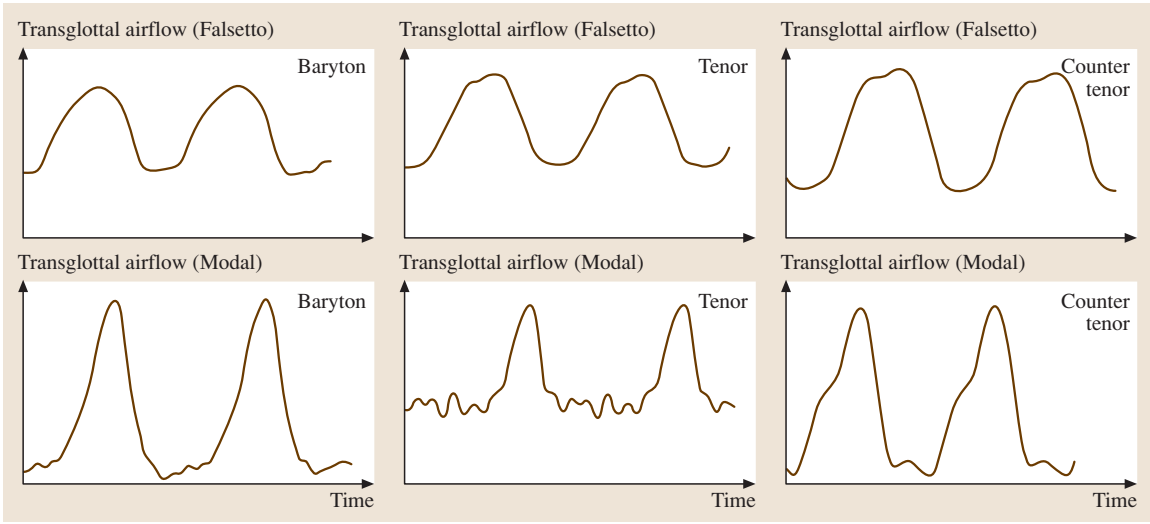


Fig. 16.16 Typical flow glottograms for falsetto and modal register in a baritone, tenor and countertenor singer, all professional. The flow scale is the same within subjects. The ripple during the closed phase in the lower middle glottogram is an artifact. In spite of the great inter-subject variability, it can be seen that the glottogram pulses are wider and more rounded in the falsetto register

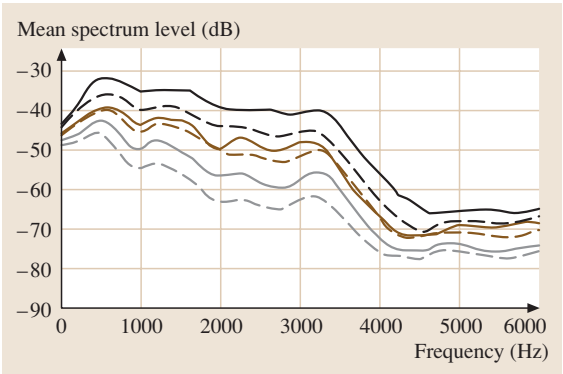


Fig. 16.17 Long-term-average spectra curves obtained from an untrained male speaker reading the same text at 6 different degrees of vocal loudness. From *top to bottom* the corresponding L_{eq} values at 0.3 m were 93 dB, 88 dB, 85 dB, 84 dB, 80 dB, and 76 dB

Variation of vocal loudness affects the spectrum slope as illustrated in Fig. 16.17, which shows long-term-average spectra (LTAS) from a male untrained voice. In the figure loudness is specified in terms of the so-called equivalent sound level L_{eq}. This is a commonly used time average of sound level, defined as

$$L_{eq} = 10 \log \frac{1}{T} \int_0^T \frac{p^2}{p_0^2} \mathrm{d}t \,,$$

where t is time and T the size of the time window. p and p_0 are the sound pressure and the reference pressure, respectively.

When vocal loudness is changed, the higher overtones change much more in sound level than the lower overtones. In the figure, a 14 dB change of the level near 600 Hz is associated with a 22 dB change near 3000 Hz, i. e., about 1.5 times the level change near 600 Hz. Similar relationships have been observed for professional singers [16.47]. In other words, the slope of the voice source spectrum decreases with increasing vocal loudness.

The physiological variable used for variation of vocal loudness is Ps. This is illustrated in the upper graph of Fig. 16.18, comparing averaged data observed in untrained female and male subjects and data obtained from professional operatic baritone singers [16.45, 46]. The relationship between the Ps and MFDR is approximately linear. It can be observed that the pressure range used by the singer is considerably wider than that used by the untrained voices. The MFDR produced with a given Ps by the untrained female and male subjects is mostly

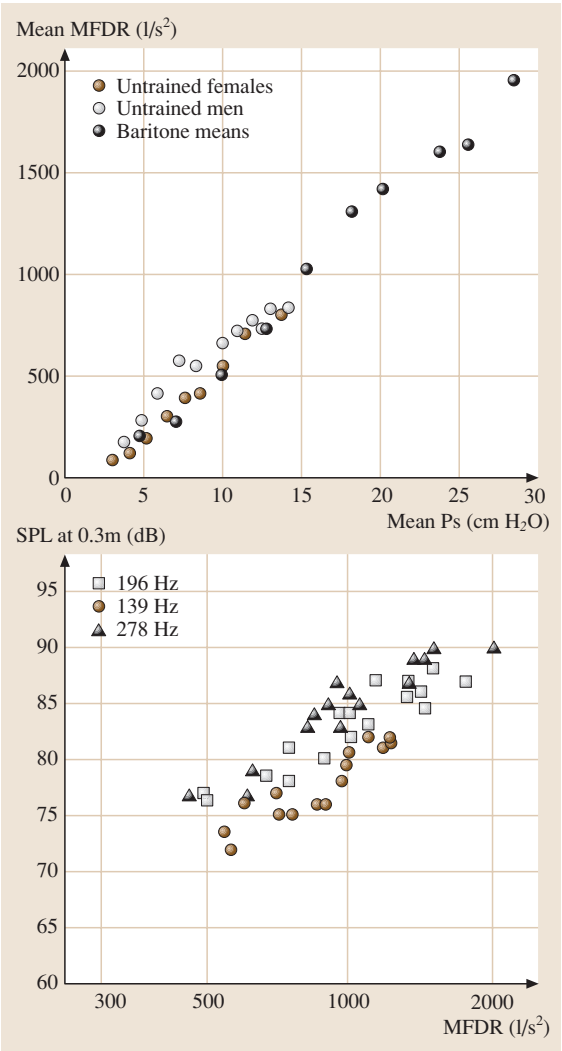


Fig. 16.18 The *top* graph shows the relationship between the mean subglottal pressure and the mean MFDR for the indicated subject groups. The *bottom* graph shows the relationship between MFDR and the SPL at 0.3 m for a professional baritone singing the vowels /a/ and /æ/ at different F0s. (After [16.45] p. 183, [16.46] p. 184)

higher than that produced by the baritones with the same pressure.This may depend on different mechanical characteristics of the vocal folds.

As we will see later, SPL depends on the strength of the excitation of the vocal tract, i. e. on MFDR. This variable, in turn, depends on Ps and F0; the higher the pressure, the greater the MFDR value and the higher the F0, the greater the MFDR. The top graph of Fig. 16.18

shows how accurately MFDR could be predicted from Ps and F0 for previously published data for untrained male and female singers and for professional baritone singers [16.45, 46]. Both Ps and F0 are linearly related to MFDR. However, the singers showed a much greater variation with F0 than the untrained voices. This difference reflected the fact that unlike the untrained subjects the singers could sing a high F0 much more softly than the untrained voices. The ability to sing high notes also softly would belong to the essential expressive skills of a singer. Recalling that an increase of Ps increases F0 by a few Hz/cm H_2O, we realize that singing high tones softly requires more forceful contraction of the pitch-raising laryngeal muscles than singing such tones loudly.

16.3 The Vocal Tract Filter

The source-filter theory, schematically illustrated in Fig. 16.19, describes vocal sound production as a three-step process: (1) generation of a steady flow of air from the lungs (DC component); (2) conversion of this airflow into a pseudo-periodically pulsating transglottal airflow (DC-to-AC conversion), referred to as the voice source; and (3) response of the vocal tract to this excitation signal (modulation of AC signal) which is characterized by the frequency curve or transfer function of the vocal tract. So far the first two stages, respiration and phonation, have been considered.

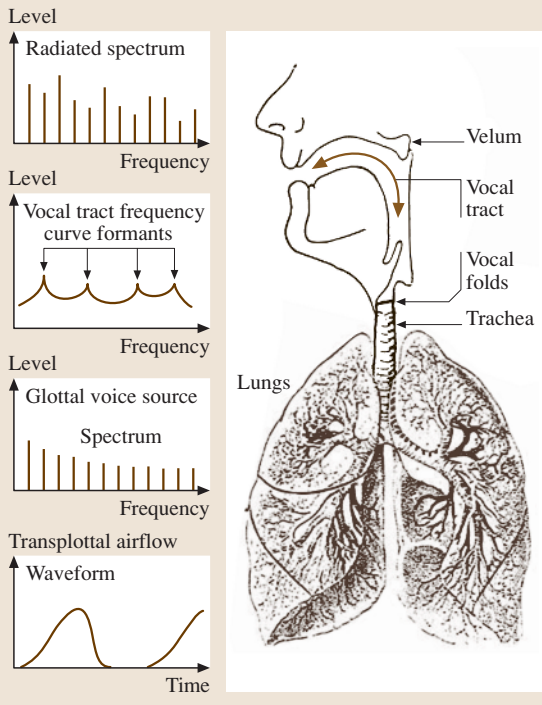


Fig. 16.19 Schematic illustration of the generation of voice sounds. The vocal fold vibrations result in a sequence of voice pulses (*bottom*) corresponding to a series of harmonic overtones, the amplitudes of which decrease monotonically with frequency (*second from bottom*). This spectrum is filtered according to the sound transfer characteristics of the vocal tract with its peaks, the formants, and the valleys between them. In the spectrum radiated from the lip opening, the formants are depicted in terms of peaks, because the partials closest to a formant frequency reach higher amplitudes than neighboring partials

In this section we will discuss the third step, viz. how the vocal tract filter, i. e. the resonance characteristics of the vocal tract, modifies, and to some extent interacts with, the glottal source and shapes the final sound output radiated from the talker's/singer's lips.

Resonance is a key feature of the filter response. The oral, pharyngeal and nasal cavities of the vocal tract form a system of resonators. During each glottal cycle the air enclosed by these cavities is set in motion by the glottal pulse, the main moment of excitation occurring during the closing of the vocal folds, more precisely at the time of the MFDR, the maximum flow declination rate (cf. the previous section on source).

The behavior of a vocal tract resonance, or *formant*, is specified both in the time and the frequency domains. For any transient excitation, the time response is an exponentially decaying cosine [16.27, p. 46]. The frequency response is a continuous amplitude-frequency spectrum with a single peak. The shape of either function is uniquely determined by two numbers (in Hz): the formant frequency F and the bandwidth B. The bandwidth quantifies the degree of damping, i. e., how fast the formant oscillation decays. Expressed as sound pressure variations, the time response is

$$p(t) = A\,\mathrm{e}^{-\pi Bt}\cos(2\pi Ft)\ . \tag{16.1}$$

For a single formant curve, the amplitude variations as a function of frequency f is given (in dB) by

$$L(f) = 20\log\frac{\left[F^2+\left(\frac{B}{2}\right)^2\right]}{\sqrt{(f-F)^2+\left(\frac{B}{2}\right)^2}\sqrt{(f+F)^2+\left(\frac{B}{2}\right)^2}}\ . \tag{16.2}$$

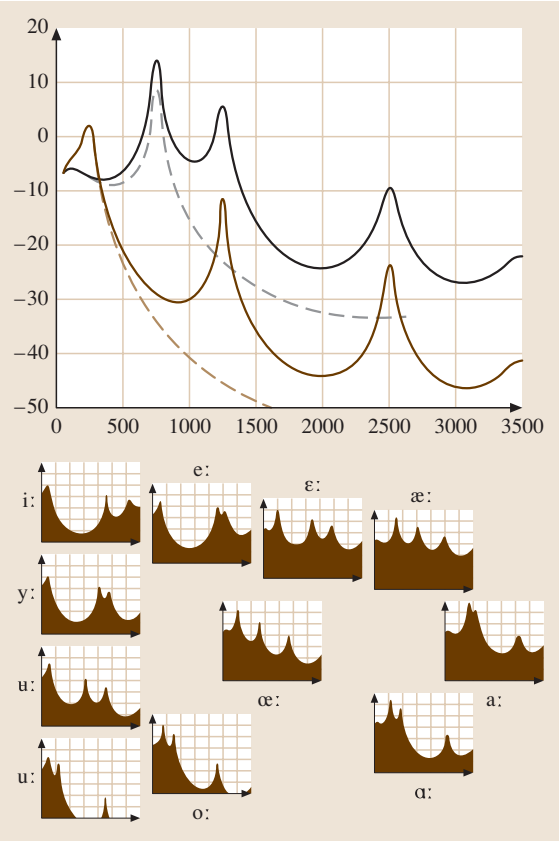


Fig. 16.20 Predictability of formant levels. Calculated spectral envelopes for a set of Swedish vowels. The *upper graph* illustrates how varying only the frequency of F1 affects the amplitudes of other formants

In the frequency domain the bandwidth is defined as the width of the formant 3 dB down from the peak, that is, at the half-power points. A large bandwidth produces a flat peak whereas a small value (less damping, reduced acoustic losses) makes the peak higher and sharper.

Figure 16.20 shows spectral envelopes for a set of Swedish vowels. They were calculated from formant frequency data [16.48, 49] in accordance with source-filter theory which assumes that a vowel spectrum can be decomposed into formants (all specified with respect to frequency and bandwidth), the spectrum of the voice source, the contribution of formants above F4 and radiation [16.27].

The individual panels are amplitude versus frequency plots. Vowels are portrayed in terms of their envelopes rather than as line spectra with harmonics. The panels are arranged in a *formant space*. From top to bottom the F2 in the panels decreases. From left to right F1 increases. The frequency scale is limited to showing the first three formant peaks.

We note that all envelopes have falling overall slopes and that the amplitudes of the formant peaks vary a great deal from vowel to vowel and depending on the relative configuration of formant frequency positions.

In acoustic phonetic specifications of vowels, it is customary to report no more than the frequencies of the first two or three formants ([16.50, p. 59], [16.51]). Experiments in speech synthesis [16.52] have indicated that this compact description is sufficient to capture the quality of steady-state vowels reasonably well. Its relative success is explained by the fact that most of the building blocks of a vowel spectrum are either predictable (bandwidths and formant amplitudes), or show only limited spectral variations (source, radiation, higher-formant correction) [16.27].

Formant bandwidths [16.28, 53] reflect acoustic losses. They depend on factors such as radiation, sound transmission through the vocal tract walls, viscosity, heat conduction, constriction size as well as the state of the glottis. For example, a more open glottis, as in a breathy voice, will markedly increase the bandwidth of the first formant.

Despite the complex interrelations among these factors, bandwidths pattern in regular ways as a function of formant frequency. Empirical formulas have been proposed [16.54] that summarize bandwidth measurements made using transient and sweep-tone excitation of the vocal tract for closed-glottis conditions [16.55, 56].

To better understand how formant levels vary let us consider the top diagram of Fig. 16.20. It compares envelopes for two vowel spectra differing only in terms of F1. It is evident that the lowering of F1 (from 750 to 250 Hz) reduces the amplitudes of F2 and F3 by about 15 dB. This effect is predicted by acoustic theory which derives the spectral envelope of an arbitrary vowel as a summation of individual formant curves on a dB scale [16.27]. Figure 16.20 makes clear that, as F1 is shifted, its contribution to the envelope is drastically changed. In this case the shift moves the entire F1 curve down in frequency and, as a result, its upper skirt (dashed line) provides less of a *lift* to the upper formants. This interplay between formant frequency and formant levels is the major determinant of the various envelope shapes seen in Fig. 16.20.

One of the main lessons of Fig. 16.20 is accordingly that, under normal conditions of a stable voice source, formant amplitudes are predictable. Another

Amazon EX1014 Page 691

important consequence of the source-filter theory is that, since knowing the formants will enable us to reconstruct the vowel's envelope, it should also make it possible to derive estimates about a vowel's overall intensity.

A vowel's intensity can be calculated from its power spectrum as

$$I = 10 \log \left[\sum (A_i)^2 \right] , \qquad (16.3)$$

where A_i is the sound pressure of the i-th harmonic. This measure tends to be dominated by the strongest partial or partials. In very soft phonation the strongest partial is generally the fundamental while in neutral and louder phonation it is the partial that lies closest to the first formant. Thus, typically all partials which are more than a few dB weaker than the strongest partial in the spectrum do not contribute appreciably to the vowel's intensity.

As suggested by the envelopes in Fig. 16.20, the strongest harmonics in vowels tend to be found in the F1 region. This implies that a vowel's intensity is primarily determined by its F1. Accordingly, in the set shown in Fig. 16.20, we would expect [iːyːɯːuː] to be least intense and [æːɑɑː] the most intense vowels. This is in good agreement with experimental observations [16.57].

The intrinsic intensity of vowels and other speech sounds has been a topic of interest to phoneticians studying the acoustic correlates of stress [16.58]. It is related to sonority, a perceptual attribute of speech sounds that tends to vary in a regular way in syllables [16.59]. We will return below to that topic in Sect. 16.5.

Let us continue to illustrate how formant levels depend on formant frequencies with an example from singing: the singer's formant. This striking spectral phenomenon is characteristic of classically trained male singers. It is illustrated in Fig. 16.21. The figure shows a spectrogram of a commercial recording of an operatic tenor voice (Jussi Björling) performing a *recitativo* from Verdi's opera Aida. Apart from the vibrato undulation of the partials, the high levels of partials in the frequency range 2200–3200 Hz are apparent. They represent the singer's formant.

The singer's formant is present in all voiced sounds as sung by operatic male singers. It was first discovered by *Bartholomew* [16.60]. It manifests itself as a high, marked peak in the long-term-average spectrum (LTAS).

A second example is presented in Fig. 16.22 showing an LTAS of the same tenor voice as in Fig. 16.21, accompanied by an orchestra of the traditional western

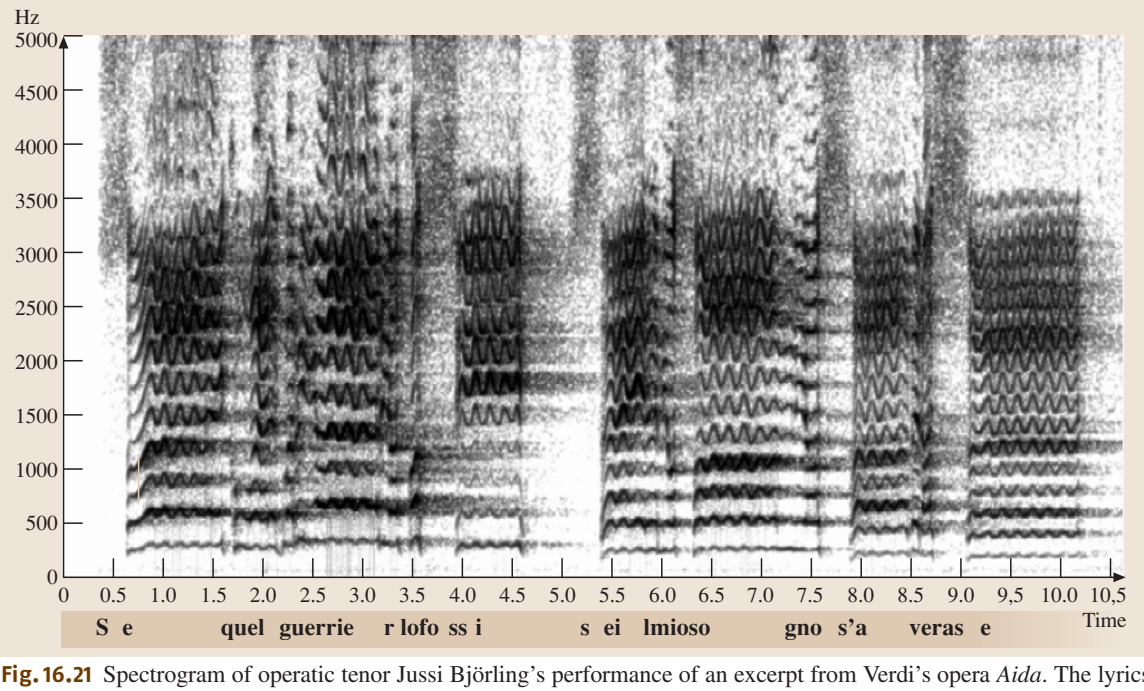


Fig. 16.21 Spectrogram of operatic tenor Jussi Björling's performance of an excerpt from Verdi's opera *Aida*. The lyrics are given below the graph

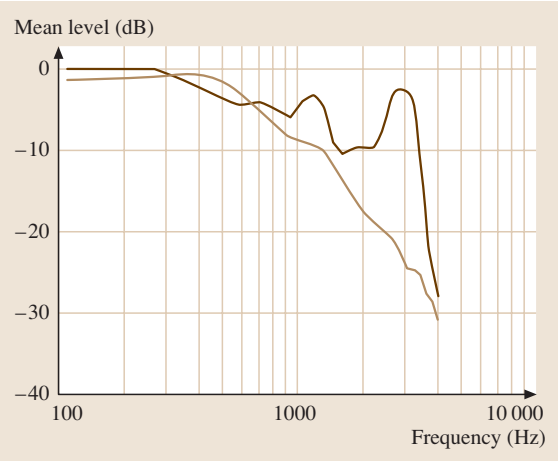


Fig. 16.22 LTAS of the sound of a symphonic orchestra with and without a singer soloist (*dark* and *light curves*). The singer's formant constitutes a major difference between the orchestra with and without the singer soloist. (After [16.61])

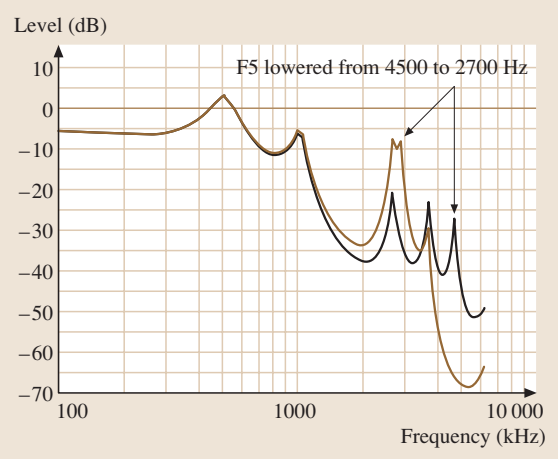


Fig. 16.23 Effect on the spectrum envelope of lowering formants F4 and F5 from 3500 Hz and 4500 Hz to 2700 Hz and 3500 Hz, respectively. The resulting gain in level at F3 is more than 10 dB

opera type. The LTAS peak caused by the singer's formant is just a few dB lower than the main peak in the low-frequency range.

In the same figure an LTAS of a classical symphony orchestra is also shown. On the average, the partials in the low-frequency range are quite strong but those above 600 Hz decrease by about 10 dB/octave with rising frequency. Even though this decrease varies depending on how loudly the orchestra is playing, this implies that the level of the orchestral accompaniment is much lower at 3000 Hz than at about 600 Hz. In other words, the orchestral sound offers the singer a rather reasonable competition in the frequency range of the singer's formant. The fact of the matter is that a voice possessing a singer's formant is much easier to hear when the orchestral accompaniment is loud than a voice lacking a singer's formant. Thus, it helps the singer's voice to be heard when the orchestral accompaniment is loud.

The singer's formant can be explained as a resonance phenomenon [16.62]. It is a product of the same rules that we invoked above to account for the formant amplitudes of vowels and for intrinsic vowel intensities. The strategy of a classically trained male singer is to shape his vocal tract so as to make F3, F4, and F5 form a tight cluster in frequency. As the frequency separations among these formants are decreased, their individual levels increase, and hence a high spectral peak is obtained between 2500 and 3000 Hz.

Figure 16.23 shows the effects on the spectrum envelope resulting from lowering F5 and F4 from 3500 Hz and 4500 Hz to 2700 Hz and 3500 Hz, respectively. The resulting increase of the level of F3 amounts to 12 dB, approximately. This means that male operatic singers produce a sound that can be heard more easily through a loud orchestral accompaniment by tuning vocal tract resonances rather than by means of producing an excessive Ps.

The acoustical situation producing the clustering of F3, F4, and F5 is obtained by acoustically mismatching the aperture of the larynx tube, also referred to as the epilaryngeal tube, with the pharynx [16.62]. This can be achieved by narrowing this aperture. Then, the larynx tube acts as a resonator with a resonance that is not much affected by the rest of the vocal tract but rather by the shape of the larynx tube. Apart from the size of the aperture, the size of the laryngeal ventricle would be influential: the larger the ventricle, the lower the larynx tube resonance. Presumably singers tune the larynx tube resonance to a frequency close to F3. The articulatory means used to establish this cavity condition seems mainly to be a lowering of the larynx, since this tends to widen both the pharynx and the laryngeal ventricle. Many singing teachers recommend students to sing with a comfortably low larynx position.

The level of the singer's formant is influenced also by the slope of the source spectrum which, in turn, depends on vocal loudness, i. e. on Ps, as mentioned. Thus, the singer's formant tends to increase by about 15 dB for a 10 dB change of the overall SPL [16.47].

Amazon EX1014 Page 693

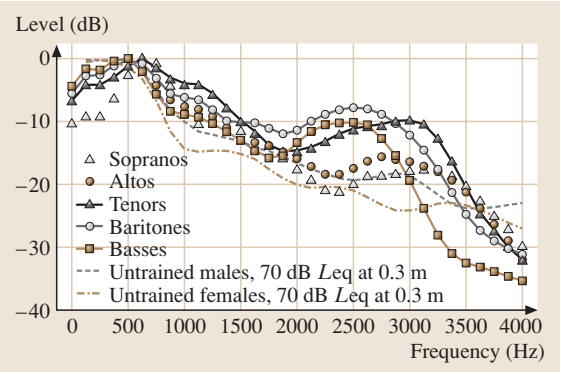


Fig. 16.24 Mean LTAS derived from commercial recordings of four representatives of each of the indicated voice classifications. The *dashed curves* show LTAS of untrained female (red) and male (blue) voices' speech with an L_{eq} of 70 dB at 0.3 m. Note that the singer's formant produces a prominent peak with a level well above the level of the untrained voices' LTAS only for the male singers

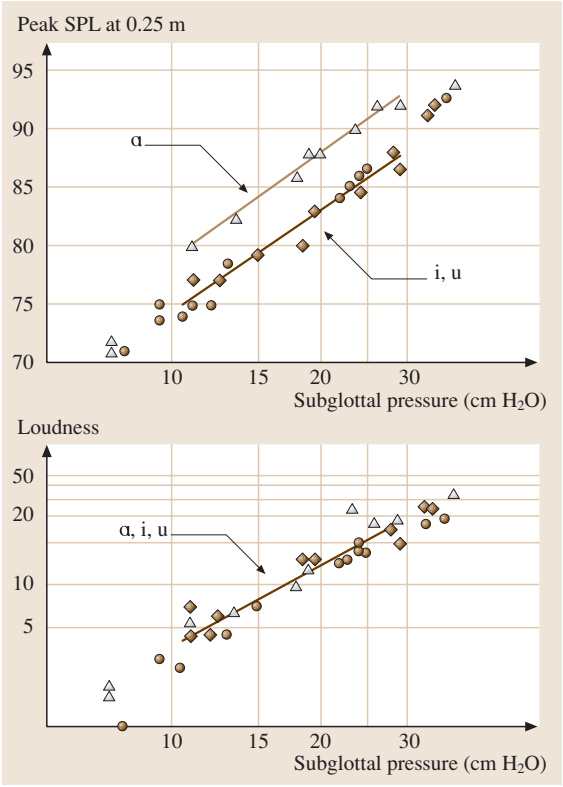


Fig. 16.25 Perceived loudness of a vowel (*bottom panel*) correlates better with subglottal pressure Ps than does SPL (*top panel*) which depends on the intrinsic intensity of the vowel. This, in turn, is determined by its formant pattern. (After [16.66])

The center frequency of the singer's formant varies slightly between voice classifications, as illustrated in the mean LTAS in Fig. 16.24 which shows mean LTAS derived from commercial recordings of singers classified as soprano, alto, tenor, baritone and bass [16.63]. Each group has four singers. The center frequency of the singer's formant for basses, baritones and tenors are about 2.4, 2.6, and 2.8 kHz, respectively. These small differences are quite relevant to the typical voice timbres of these classifications [16.64]. Their origin is likely to be vocal tract length differences, basses tending to have longer vocal tracts than baritones who in turn have longer vocal tracts than tenors [16.65]. On the other hand, substantial variation in the center frequency of the singer's formant occurs also within classifications.

Also shown for comparison in Fig. 16.24 is a mean LTAS of untrained voices reading at an L_{eq} of 70 dB at 0.3 m distance. The singer's formant produces a marked peak some 15 dB above the LTAS of the untrained voices for the male singers. The female singers, on the other hand, do not show any comparable peak in this frequency range. This implies that female singers do not have a singer's formant [16.67–69].

Female operatic singers' lack of a singer's formant is not surprising, given the fact that: (1) they sing at high F0, i. e. have widely spaced spectrum partials, and (2) the F3, F4, F5 cluster that produces a singer's formant is rather narrow in frequency. The latter means that a narrow formant cluster will be hit by a partial only in some tones of a scale while in other scale tones there will be no partial in the formant cluster. This would lead to salient and quasi-random variation of voice timbre between scale tones.

The singer's formant is a characteristic of classically trained male singers. It is not found in nonclassical singing, e.g. in pop or musical theater singing, where audibility is the responsibility of the sound engineer rather than of the singer. Likewise, choir singers generally do not possess a singer's formant.

From the evidence in Fig. 16.18 we concluded that the logarithm of Ps does a good job of predicting the SPL over a large range of vocal intensities. This was shown for untrained male and female speakers as well as a professional baritone singer. In the next few paragraphs we will look at how vowel identity affects that prediction and how listeners perceive loudness in the presence of vowel-dependent variations in SPL at constant Ps.

Amazon EX1014 Page 694

This topic was addressed by Ladefoged [16.66] who measured the peak SPL and the peak Ps in 12 repetitions of *bee*, *bay*, *bar*, *bore* and *boo* spoken by a British talker at varying degrees of loudness. SPL values were plotted against log(Ps) for all tokens. Straight lines could readily be fitted to the vowels individually. The vowels of *bay* and *bore* tended to have SPL values in between those of *bar* and *bee*/*boo*. The left half of Fig. 16.25 replots the original data for [ɑ] and [i]/[u] pooled. The [ɑ] line is higher by 5–6 dB as it should be in view of F1 being higher in [ɑ] than in [i] and [u] (cf. preceding discussion).

In a second experiment listeners were asked to judge the loudness of the test words. Each item was presented after a carrier phrase: *Compare the words: bar and* __. The word bar served as reference. The subjects were instructed to compare the test syllable and the reference in terms of loudness, to give the value of 10 to the reference and another relative number to the test word. The analysis of the responses in terms of SPL indicated that, for a given SPL, it was consistently the case that *bee* and *boo* tended to be judged as louder than *bar*. On the other hand, there were instances of *bee* and *bar* with similar Ps that were judged to be equally loud. This effect stood out clearly when loudness judgements were plotted against log(Ps) as in the bottom half of Fig. 16.25. Ladefoged concludes that " ... in the case of speech sounds, loudness is directly related to the physiological effort" – in other words, Ps – rather than the SPL as for many other sounds.

Speech Sounds with Noise and Transient Excitation

A comprehensive quantitative treatment of the mechanisms of noise production in speech sounds is found in *Stevens* [16.28, pp. 100–121]. This work also provides detailed analyses of how the noise source and vocal tract filtering interact in shaping the bursts of stops and the spectral characteristics of voiced and voiceless fricatives. While the normal source mechanism for vowels always involves the glottis, noise generation may take place not only at the glottis but at a wide range of locations along the vocal tract. A useful rule of thumb for articulations excited by noise is that the output will spectrally be dominated by the cavity in front of the noise source. This also holds true for sounds with transient excitation such as stop releases and click sounds [16.28, Chapts. 7,8].

While most of the preceding remarks were devoted to the acoustics of vowels, we should stress that the source-filter theory applies with equal force to the production of consonants.

16.4 Articulatory Processes, Vowels and Consonants

X-ray films of normal speech movements reveal a highly dynamic process. Articulatory activity comes across as a complex flow of rapid, parallel lip, tongue and other movements which shows few, if any, steady states. Although the speaker may be saying no more than a few simple syllables, one nonetheless has the impression of a virtuoso performance of a *polyphonic* motor score. As these events unfold, they are instantly reflected in the acoustic output. The articulatory movements modify the geometry of the vocal tract. Accordingly, the filter (transfer function) undergoes continual change and, as a result, so do the output formant frequencies.

Quantitative modeling is a powerful tool for investigating the speech production process. It has been successfully applied to study the relations between formant and cavities. Significant insights into this mapping have been obtained by representing a given vocal tract configuration as an *area function* – i. e., a series of cylindrical cross sections of variable lengths and cross-sectional areas – and by simulating the effects of changes in the area function on the formant frequencies [16.27].

In the past, pursuing this approach, investigators have used lateral X-ray images – similar to the magnetic

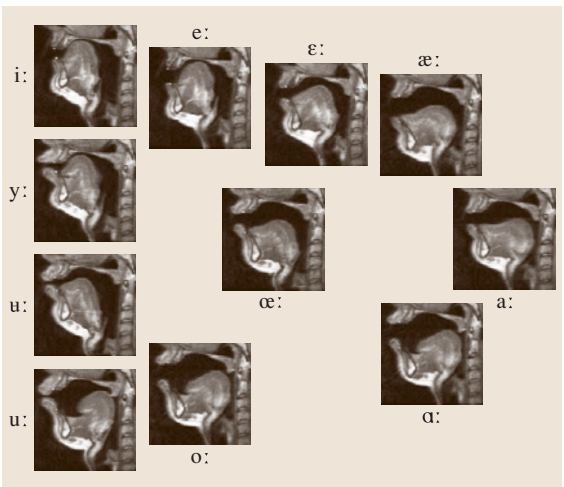


Fig. 16.26 Magnetic resonance images of the Swedish vowels of Fig. 16.20

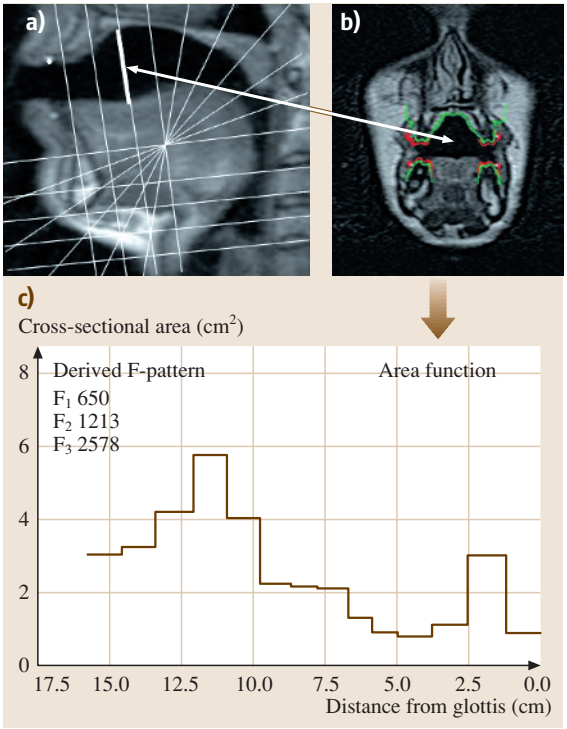


Fig. 16.27a–c From articulatory profile to acoustic output. Deriving the formant patterns of a given vowel articulation involves **(a)** representing the vocal tract profile in terms of the cross-distances from the glottis to the lips; **(b)** converting this specification into a cross-sectional area function; and **(c)** calculating the formant pattern from this area function. To go from **(a)** to **(b)**, profile data (*top left*) need to be supplemented by area measurements obtained from transversal (coronal and axial) images of the cross sections

resonance imaging (MRI) pictures in Fig. 16.26 – to trace the outlines of the acoustically relevant articulatory structures. To make estimates of cross-sectional areas along the vocal tract such lateral profiles need to be supplemented with information on the transverse geometry of the cross sections, e.g., from casts of the vocal tract [16.70] and tomographic pictures [16.27, 71].

More currently, magnetic resonance imaging methods have become available, making it possible to obtain three-dimensional (3-D) data on sustained steady articulations [16.72–76]. Figure 16.26 presents MRI images taken in the mid-sagittal plane of a male subject during steady-state productions of a set of Swedish vowels. These data were collected in connection with work on APEX, an articulatory model developed for studying the acoustic consequences of articulatory movements, e.g., the lips, the tongue and the jaw [16.77–81].

Figure 16.27 highlights some of the steps involved in deriving the formant pattern from lateral articulatory profiles such those of Fig. 16.26. The top-left picture is the profile for the subject's [a]. The white lines indicate the coronal, coronal oblique and axial planes where ad-

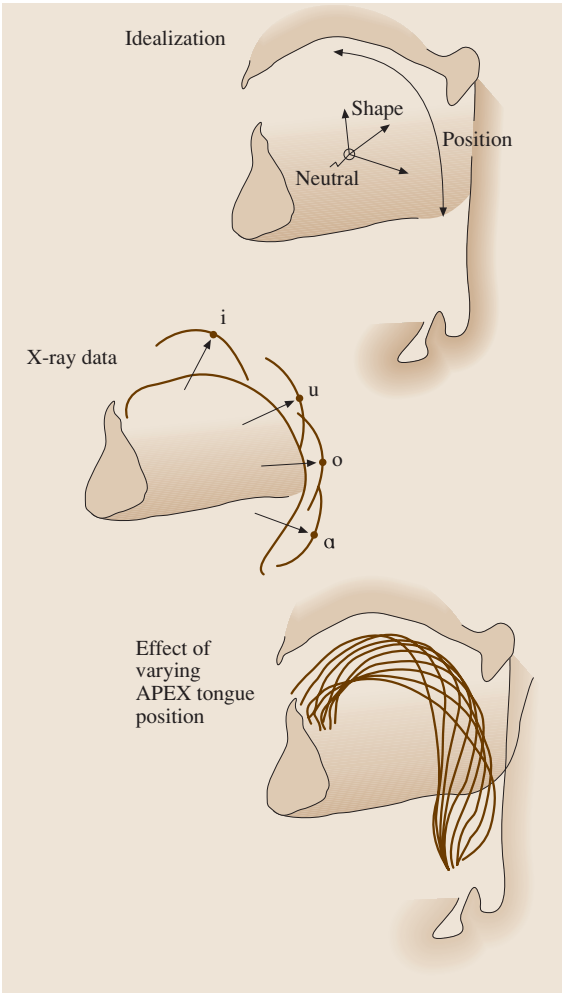


Fig. 16.28 The top drawing shows superimposed observed tongue shapes for [i], [u], [o] and [ɑ] differing in the place of the main constriction. Further analyses of similar data suggest that, for vowels, two main dimensions are used: The anterior–posterior location of the tongue body and its displacement from neutral which controls the degree of vocal tract constriction (*lower right*). These parameters are implemented numerically to produce vowels in the APEX articulatory model

Amazon EX1014 Page 696

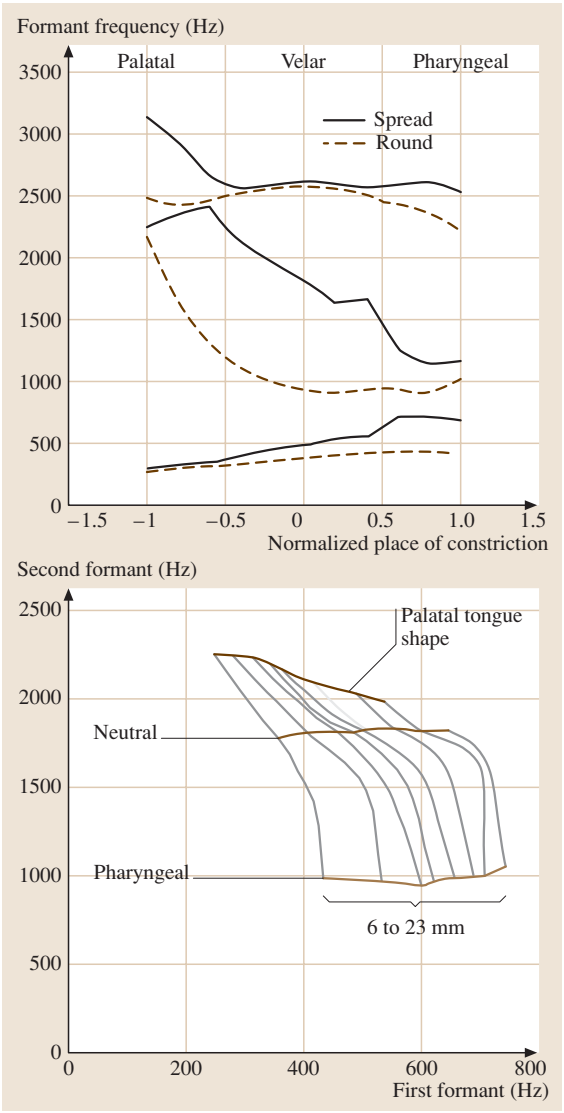


Fig. 16.29 Acoustic consequences of tongue body, jaw and lip movements as examined in APEX simulations. The tongue moves continuously between palatal, velar and pharyngeal locations while maintaining a minimum constriction area A_{min} of 0.25 cm, a 7 mm jaw opening and a fixed larynx height. The *dashed* and *solid lines* refer to rounded and spread conditions. The *lower diagram* illustrates the effect of varying the jaw. The *thin lines* show tongue movement at various degrees of jaw opening (6, 7, 8, 9, 12, 15, 19 and 23 mm). The *bold lines* pertain to fixed palatal, neutral or pharyngeal tongue contours

ditional MR images were taken to get information on the cross sections in the transverse dimension.

The location of the transverse cut is indicated by the bold white line segment in the lateral profile. In the coronal section to the right the airway of the vocal tract is the dark area at the center. Below it, the wavy surface of the tongue is evident. Since the teeth lack the density of hydrogen nuclei needed to show up on MR images, their intersection contours (green) were reconstructed from casts and added computationally to the image. The red lines indicate the outline of a dental plate custom-fitted to the subject's upper and lower teeth and designed to carry a contrast agent used to provide reference landmarks (tiny white dots) [16.82].

At any given point in the vocal tract, it would appear that changes in cross-sectional area depend primarily on the midsagittal distance from the tongue surface to the vocal tract wall. An empirically adequate approach to capturing distance-to-area relations is to use power functions of the form $A_x = \alpha d_x^{\beta}$, where d_x is the mid-sagittal cross distance at a location x in the vocal tract and α and β are constants whose values tend to be different in different regions and depend on the geometry of the vocal tract walls [16.24,70,72]. The cross-sectional area function can be obtained by applying a set of d-to-A rules of this type to cross distances measured along the vocal tract perpendicular to the midline. The final step consists in calculating the acoustic resonance frequencies of the derived area function.

In producing a vowel and the vocal tract shape appropriate for it, what are the main articulatory parameters that the speaker has to control? The APEX model [16.80] assumes that the significant information about the vowel articulations in Fig. 16.26 is the following. All the vowels exhibit tongue shapes with a single constriction. They differ with respect to how narrow this constriction is and where it is located. In other words, the vowels appear to be produced with control of two degrees of freedom: the palatal–pharyngeal dimension (*position*) and tongue height, or, in APEX terminology, *displacement* from neutral. This interpretation is presented in stylized form in Fig. 16.28 together with APEX tongue shapes sampled along the palatal–pharyngeal continuum. The choice of tongue parameters parallels the degree and place of constriction in the area-function models of *Stevens* and *House* [16.83] and *Fant* [16.27].

A useful application of articulatory models is that they can be set up to change a certain variable while keeping others constant [16.84]. Clearly, asking a human subject to follow such an instruction does not create an easily controlled task.

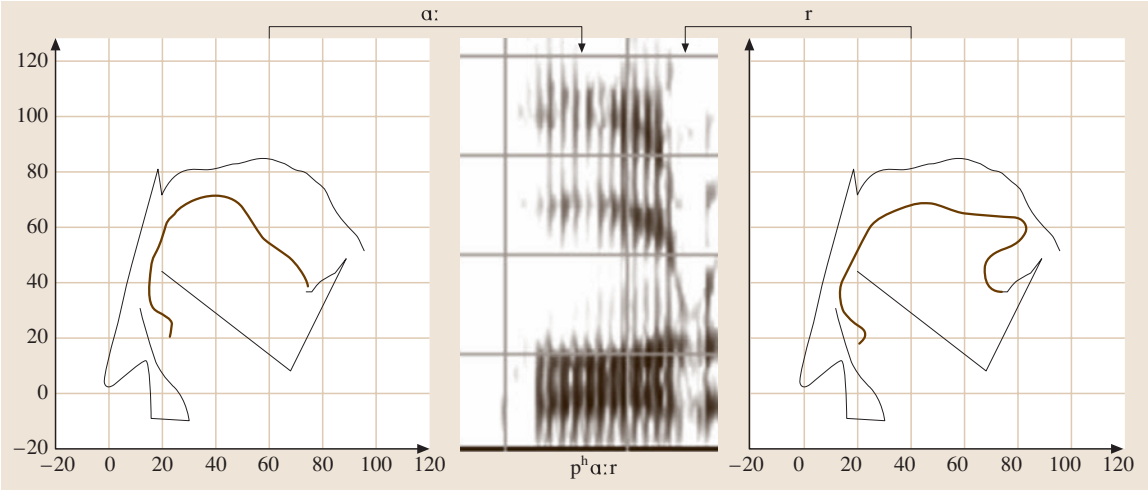


Fig. 16.30 Role of the tongue tip and blade in tuning F3. As the tongue is raised to produce the retroflex configuration required for the Swedish [r] sound an acoustically significant volume is created under the tongue blade. Simulations using APEX indicate that this cavity is responsible for the considerable lowering of the third formant associated with the [r]

Figure 16.29 plots results from APEX simulation experiments. It demonstrates how F1, F2 and F3 vary in response to the tongue moving continuously between palatal, velar and pharyngeal locations and maintaining a minimum constriction area A_{min} of 0.25 cm^2, a 7 mm jaw opening and keeping the larynx height constant. The dashed and solid lines refer to rounded and spread conditions. The calculations included a correction for the impedance of the vocal tract walls ([16.54], [16.28, p. 158]).

It is seen that the tongue movement has strong effect on F2. As the constriction becomes more posterior F2 decreases and F1 rises. In general, rounding lowers formants by varying degrees that depends on the formant-cavity relations that apply to the articulation in question. However, for the most palatal position in Fig. 16.29 – an articulation similar to an [i] or an [y], it has little effect on F2 whereas F3 is affected strongly.

The F2 versus F1 plot of the lower diagram of Fig. 16.29 was drawn to illustrate the effect of varying the jaw. The thin lines show how, at various degrees of jaw opening (6, 7, 8, 9, 12, 15, 19 and 23 mm), the tongue moves between palatal and pharyngeal constrictions by way of the neutral tongue shape. The bold lines pertain to fixed palatal, neutral or pharyngeal tongue contours. We note that increasing the jaw opening while keeping the tongue configuration constant shifts F1 upward.

Figure 16.30 exemplifies the role of the tongue tip and blade in tuning F3. The data come from an X-ray study with synchronized sound [16.79,85]. At the center a spectrogram of the syllable "*par*" [pʰɑːr]. Perhaps the most striking feature of the formant pattern is the extensive lowering of F3 and F4 into the [r].

The articulatory correlates are illustrated in the tracings. As we compare the profiles for [ɑː] and [r], we see that the major change is the raising of the tongue blade for [r] and the emergence of a significant cavity in front of, and below, the tongue blade. Simulations us-

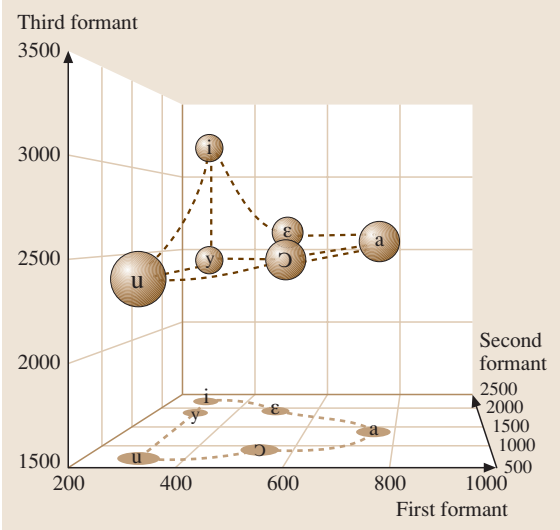


Fig. 16.31 The acoustic vowel space: *possible* vowel formant combinations according to the APEX model

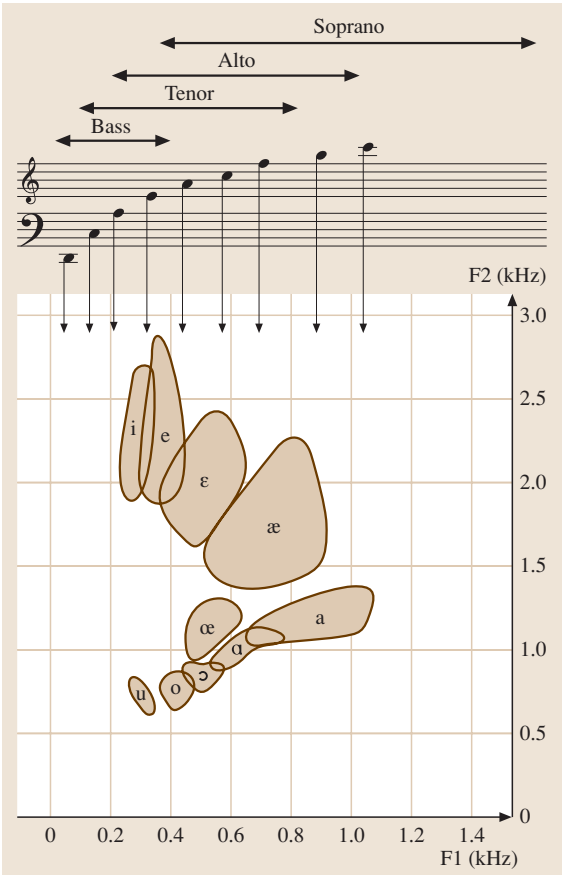


Fig. 16.32 Schematic illustration of the variation ranges of F1 and F2 for various vowels as pronounced by female and male adults. Above the graph F1 is given in musical score notation together with typical F0 ranges for the indicated voice classifications

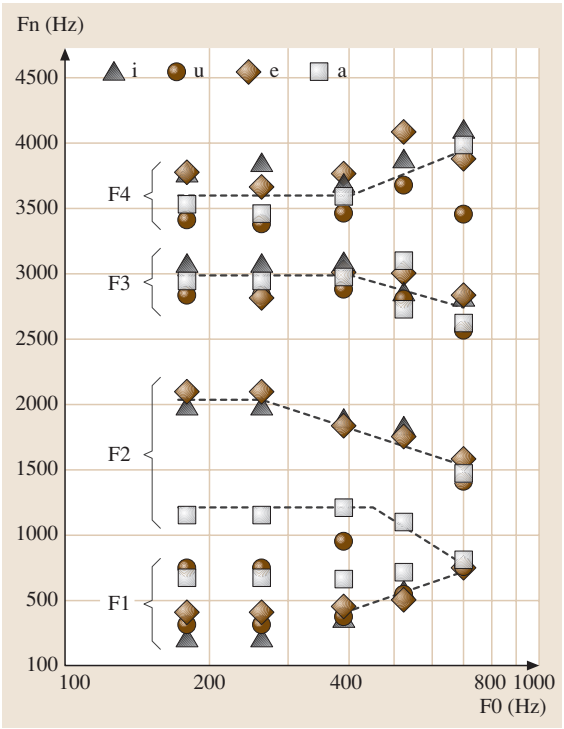


Fig. 16.33 Values of F1, F2, F3, and F4 for the indicated vowels in a professional soprano as function of F0. F1 and F3 are represented by open symbols and F2 and F4 by filled symbols. The values for F0 ≈180 Hz were obtained when the singer sustained the vowels in a speech mode. (After *Sundberg* [16.87])

ing APEX and other models [16.28,86] indicate that this subapical cavity is responsible for the drastic lowering of F3.

The curves of Fig. 16.30 are qualitatively consistent with the nomograms published for three-parameter area-function models, e.g., Fant [16.27]. An advantage of more-realistic physiological models is that the relationships between articulatory parameters and formant patterns become more transparent. We can summarize observations about APEX and the articulation-to-formants mapping as:

- F1 is controlled by the jaw in a direct way;
- F2 is significantly changed by front–back movement of the tongue;
- F3 is influenced by the action of the tongue blade.

Another way of summarizing the acoustic properties of a speech production model is to translate all of its articulatory capabilities into a formant space. By definition that contains all the formant patterns that the model is capable of producing (and specifying articulatorily). Suppose that a model uses n independent parameters and each parameter is quantized into a certain number of steps. By forming all *legal* combinations of those parametric values and deriving their vocal tract shapes and formant patterns, we obtain the data needed to represent that space graphically.

The APEX vowel space is shown in 3-D in Fig. 16.31. The x-axis is F1. The depth dimension is F2 and the vertical axis is F3. Smooth lines were drawn to enclose individual data points (omitted for clarity).

A cloud-like structure emerges. Its projection on the F2/F1 floor plane takes the form of the familiar trian-

Amazon EX1014 Page 699

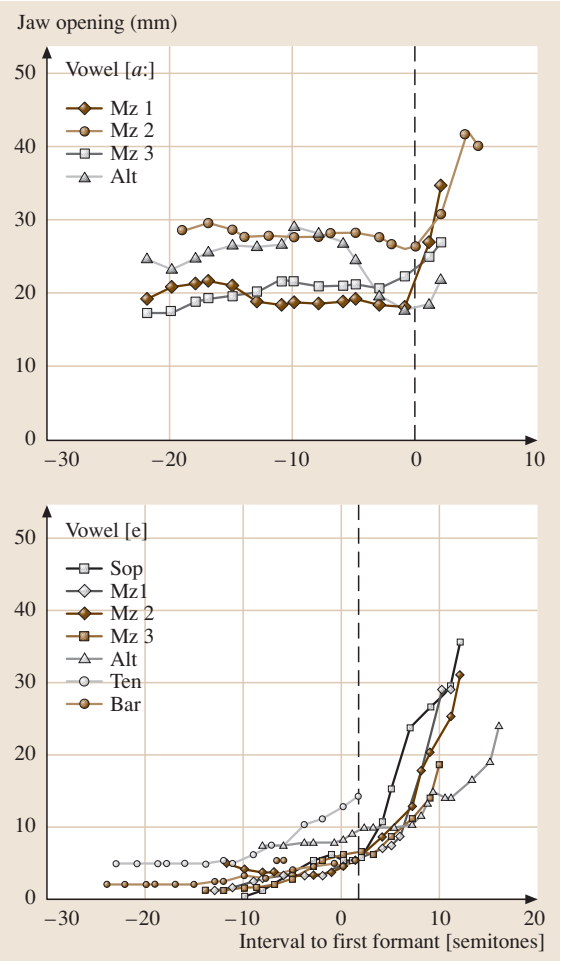


Fig. 16.34 Singers' jaw opening in the vowels [ɑ:] and [e] plotted as functions of the distance in semitones between F0 and the individual singer's normal F1 value for these vowels. The singers belonged to different classifications: Sop = soprano, Mz = mezzosoprano, Alt = alto, Ten = tenor, Bar = baritone. Symbols refer to subjects. (After *Sundberg* and *Skoog* [16.88])

gular pattern appears with [i], [a] and [u] at the corners. Adding F3 along the vertical axis offers additional room for vowel timber variations especially in the [i/y] region.

Figure 16.32 shows typical values for F1 and F2 for various vowels. At the top of the diagram, F1 is given also on the musical staff. The graph thus demonstrates that the fundamental frequency in singing is often higher than the normal value of F1. For example, the first formant of [i:] and [u:] is about 250 Hz (close to the pitch of C4), which certainly is a quite low note for a soprano.

Theory predicts that the vocal fold vibration will be greatly disturbed when F0 = F1 [16.89], and singers seem to avoid allowing F0 to pass F1 [16.87, 90]. Instead they raise F1 to a frequency slightly higher than F0. Some formant frequency values for a soprano singer are shown in Fig. 16.33. The results can be idealized in terms of lines, also shown in the figure, relating F1, F2, F3, and F4 to F0. Also shown are the subject's formant frequencies in speech. The main principle seems to be as follows. As long as fundamental frequency is lower than the normal value of the vowel's first formant frequency, this formant frequency is used. At higher pitches, the first formant is raised to a value somewhat higher than the fundamental frequency. In this way, the singer avoids having the fundamental frequency exceed the first formant frequency. With rising fundamental frequency, F2 of front vowels is lowered, while F2 of back vowels is raised to a frequency just above the second spectrum partial; F3 is lowered, and F4 is raised.

A commonly used articulatory trick for achieving the increase of F1 with F0 is a widening of the jaw opening [16.85]. The graphs of Fig. 16.34 show the jaw opening of professional singers as a function of the frequency separation in semitones between F0 and the F1 value that the singer used at low pitches. The graph referring to the vowel [ɑ:] shows that most singers began to widen their jaw opening when the F0 was about four semitones below the normal F1 value. The lower graph of Fig. 16.34 shows the corresponding data for the vowel [e]. For this vowel, most subjects started to widen the jaw opening when the fundamental was about four semitones above the normal value of F1. It is likely that below this pitch singers increase the first formant by other articulatory means than the jaw opening. A plausible candidate in front vowels is the degree of vocal tract constriction; a reduced constriction increases the first formant. Many singing teachers recommend their students to *release the jaw* or to *give space to the tone*; it seems likely that the acoustic target of these recommendations is to raise F1.

There are also other articulators that can be recruited for the purpose of raising F1. One is the lip opening. By retracting the mouth corners, the vocal tract is shortened; and hence the frequencies of the formants will increase. The vocal tract can also be shortened by raising the larynx, and some professional female singers take advantage of this tool when singing at high pitches.

This principle of tuning formant frequencies depending on F0 has been found in all singers who encounter the situation that the normal value of the first formant is lower than their highest pitches. In fact, all singers except basses encounter this situation at least for some

Amazon EX1014 Page 700

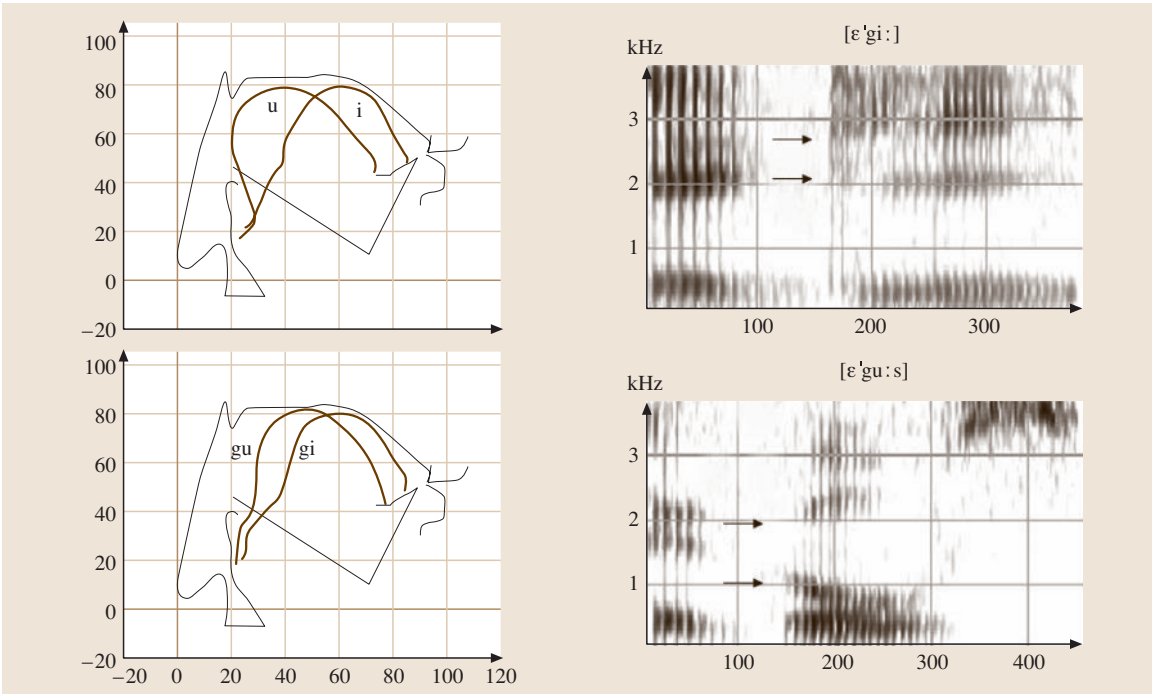


Fig. 16.35 The universal phenomenon of coarticulation illustrated with articulatory and acoustic data on Swedish [g]. The influence of the following vowel extends throughout the articulation of the stop closure. There is no point in time at which a *pure* (context-free) sample of the [g] could be obtained

vowels sung at high pitches. The benefit of these arrangements of the formant frequencies is an enormous increase of sound level, gained by sheer resonance.

Vowel quality is determined mainly by F1 and F2, as mentioned. Therefore, one would expect drastic consequences regarding vowel intelligibility in these cases. However, the vowel quality of sustained vowels seems to survive these pitch-dependent formant frequency changes surprisingly well, except when F0 exceeds the pitch of F5 (about 700 Hz). Above that frequency no formant frequency combination seems to help, and below it, the vowel quality would not be better if normal formant frequencies were chosen. The amount of text intelligibility which occurs at very high pitches relies almost exclusively on the consonants surrounding the vowel. Thus, facing the choice between inaudible tones with normal formant frequencies or audible tones with strange vowel quality, singers probably make a wise choice.

One of the major challenges both for applied and theoretical speech research is the great variability of the speech signal. Consider a given utterance as pronounced by speakers of the same dialect. A few moments' reflection will convince us that this utterance is certain to come in a large variety of physical shapes. Some variations reflect the speaker's age, gender, vocal anatomy as well as emotional and physiological state. Others are stylistic and situational as exemplified by speaking clearly in noisy environments, speaking formally in public, addressing large audiences, chatting with a close friend or talking to oneself while trying to solve a problem. Style and situation make significant contributions to the variety of acoustic waveforms that instantiate what we linguistically judge as the *same utterance*.

In phonetic experimentation investigators aim at keeping all of these stylistic and situational factors constant. This goal tends to limit the scope of the research to a style labeled *laboratory speech*, which consists of test items that are chosen by the experimenter and are typically read from a list. Despite the marked focus on this type of material, laboratory speech nonetheless presents several variability puzzles. We will mention two: segmental interaction, or *coarticulation*, and prosodic modulation. Both exemplify the ubiquitous context dependence of phonetic segments. First a few remarks on coarticulation.

Amazon EX1014 Page 701

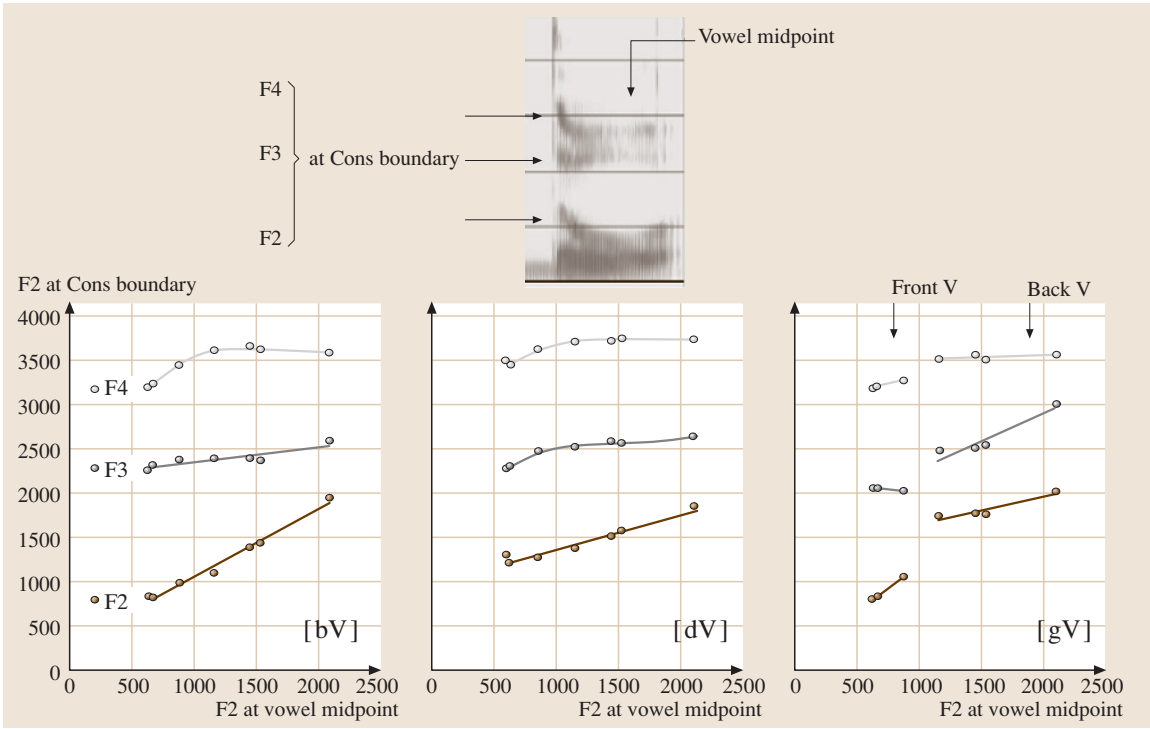


Fig. 16.36 Coarticulation in [bV], [dV] and [gV] as reflected by measurements of F2, F3 and F4 onsets plotted against an acoustic correlate of front back tongue movement, viz., F2 at the vowel midpoint

In Fig. 16.35 the phoneme [g] occurs in two words taken from an X-ray film [16.79, 85] of a Swedish speaker: [ɛˈgiː] and [ɛˈguːs]. In these words the first three phonemes correspond to the first three segments on the spectrogram: an initial [ɛ] segment, the [g] stop gap and then the final vowel. So far, so good. However, if we were to try to draw a vertical line on the spectrogram to mark the point in time where [ɛ] ends and [g] begins, or where [g] ends and the final vowel begins, we would soon realize that we have an impossible task. We would perhaps be able to detect formant movements in the [ɛ] segment indicating that articulatory activity towards the [g] had been initiated. However, that point in time occurs during the acoustic [ɛ] segment. Similarly, we could identify the endpoint of the formant transitions following [g] but that event occurs when the next segment, the final vowel, is already under way.

What we arrive at here is the classical conclusion that strings of phonemes are not organized as *beads on a necklace* [16.91–93]. The acoustic correlates of phonemes, the acoustic segments, are produced according to a motor schema that requires parallel activity in several articulatory channels and weaves the sequence of phonemes into a smooth fabric of overlapping movements. We are talking about *coarticulation*, the overlap of articulatory gestures in space and time.

Not only does this universal of motor organization give rise to the segmentation problem, i. e., make it impossible to chop up the time scale of the speech wave into phoneme-sized chunks, it creates another dilemma known as the *invariance issue*. We can exemplify it by referring to Fig. 16.35 again and the arrows indicating the frequencies of F2 and F3 at the moment of [g] release. With [iː] following they are high in frequency. Next to [uː] they are lower. What acoustic attributes define the [g] phoneme? How do we specify the [g] phoneme acoustically in a context-independent way?

The answer is that, because of coarticulation, we cannot provide an acoustic definition that is context independent. There is no such thing as an acoustic *pure sample* of a phoneme.

Articulatory observations confirm this account. There is no point in time where the shape of the tongue shows zero influence from the surrounding vowels. The tracings at the top left of Fig. 16.35 show the tongue con-

Amazon EX1014 Page 702

tours for [iː] and [uː] sampled at the vowel midpoints. The bottom left profiles show the tongue shapes at the [g] release. The effect of the following vowel is readily apparent.

So where do we look for the invariant phonetic correlates for [g]? Work on articulatory modeling [16.49, 94] indicates that, if there is such a thing as a single context-free target underlying the surface variants of [g], it is likely to occur at a deeper level of speech production located before the motor commands for [g] closure and for the surrounding vowels blend.

This picture of speech production raises a number of questions about speech perception that have been addressed by a large body of experimental work [16.95, 96] but which will not be reviewed here.

It should be remarked though that, the segmentation and invariance issues notwithstanding, the context sensitivity of phonetic segments is systematic. As an illustration of that point Fig. 16.36 is presented. It shows average data on formant transitions that come from the Swedish speaker of Fig. 16.36 and Figs. 16.26–16.28. The measurements are from repetitions of CV test words in which the consonants were [b], [d] or [g] and were combined with [iː] [ɛː] [æː] [a] [ɑː] [oː] and [uː]. Formant transition onsets for F2, F3 and F4 are plotted against F2 midpoints for the vowels.

If the consonants are coarticulated with the vowels following, we would expect consonant onset patterns to co-vary with the vowel formant patterns. As shown by Fig. 16.36 that is also what we find. Recall that, in the section on articulatory modeling, we demonstrated that F2 correlates strongly with the front–back movement of the tongue. This implies that, in an indirect way, the *x*-axis labeled 'F2 at vowel midpoint' can be said to range from back to front. The same reasoning applies to F2 onsets.

Figure 16.36 shows that the relationship between F2 onsets and F2 at vowel midpoint is linear for bV and dV. For gV, the data points break up into back (low F2) and front (high F2) groups. These straight lines – known as *locus equations* [16.97] – have received considerable attention since they provide a compact way of quantifying coarticulation. Data are available for several languages showing robustly that slopes and intercepts vary in systematic ways with places of articulation.

Furthermore, we see from Fig. 16.36 that lawful patterns are obtained also for F3 and F4 onsets. This makes sense if we assume that vocal tract cavities are not completely uncoupled and that hence, all formants – not only F2 – are to some extent influenced by where along the front–back dimension the vowel is articulated.

16.5 The Syllable

A central unit in both speech and singing is the syllable. It resembles the phoneme in that it is hard to define but it can be described in a number of ways.

Linguists characterize it in terms of how vowels and consonants pattern within it. The central portion, the nucleus, is normally a vowel. One or more consonants can precede and/or follow forming the onset and the coda respectively. The vowel/nucleus is always there; the onset and coda are optional.

Languages vary with respect to the way they combine consonants and vowels into syllables. Most of them favor a frame with only two slots: the CV syllable. Others allow more-elaborated syllable structures with up to three consonants initially and the mirror image in syllable final position. If there is also a length distinction in the vowel and/or consonant system, syllables frames can become quite complex. A rich pattern with consonant clusters and phonological length usually implies that the language has a strong contrast between stressed and unstressed syllables.

In languages that allow consonant sequences, there is a universal tendency for the segments to be serially ordered on an articulatory continuum with the consonants compatible with the vowel's greater jaw opening occurring next to the vowel, e.g., [l] and [r], while those less compatible, e.g. [s], are recruited at the syllable margins [16.98, 99]. In keeping with this observation, English and other languages use [spr] as an initial, but not final, cluster. The reverse sequence [rps] occurs in final, but not initial, position, cf. *sprawl* and *harps*. Traditionally and currently, this trend is explained in terms of an auditory attribute of speech sounds, sonority. The *sonority principle* [16.59] states that, as the most *sonorous* segments, vowels take the central nucleus position of the syllable and that the *sonority* of the surrounding consonants must decrease to the left and to the right starting from the vowel. Recalling that the degree of articulatory opening affects F1 which in turn affects sound intensity, we realize that these articulatory and auditory accounts are not incompatible. However,

Amazon EX1014 Page 703

the reason for the syllabic variations in sonority is articulatory: the tendency for syllables to alternate close and open articulations in a cyclical manner.

The syllable is also illuminated by a developmental perspective. An important milestone of normal speech acquisition is *canonical babbling*. This type of vocalization makes its appearance sometime between 6–10 months. It consists of sequences of CV-like events, e.g., [dædæ], [baba] [16.101–105]. The phonetic output of deaf infants differs from canonical babbling both quantitatively and quantitatively [16.106–108], suggesting that auditory input from the ambient language is a prerequisite for canonical babbling [16.109–112]. What babbling shares with adult speech is its "syllabic" organization, that is, the alternation of open and close articulations in which jaw movement is a major component [16.113].

As mentioned, the regular repetition of open-close vocal tract states gives rise to an amplitude modulation of the speech waveform. Vowels tend to show the highest amplitudes contrasting with the surrounding consonants which have various degrees of constriction and hence more reduced amplitudes. At the acoustic boundary between a consonant and a vowel, there is often an abrupt rise in the amplitude envelope of the waveform.

When a Fourier analysis is performed on the waveform envelope, a spectrum with primarily low, sub-audio frequency components is obtained. This is to be expected given the fact that amplitude envelopes vary slowly as a function of time. This representation is known as the modulation spectrum [16.114]. It reflects recurring events such as the amplitude changes at consonant-vowel boundaries. It provides an approximate record of the rhythmic pulsating stream of stressed and unstressed syllables.

The time envelope may at first appear to be a rather crude attribute of the signal. However, its perceptual importance should not be underestimated. Room acoustics and noise distort speech by modifying and destroying its modulation spectrum. The modulation transfer function was proposed by Houtgast and Steeneken as a measure of the effect of the auditorium on the speech signal and as a basis for an index, the speech transmission index (STI), used to predict speech intelligibility under different types of reverberation and noise. The success of this approach tells us that the modulation spectrum, and hence the waveform envelope, contains information that is crucial for robust speech perception [16.115]. Experimental manipulation of the temporal envelope has been performed by Drullman et al. [16.116, 117] whose work reinforces the conclusions reached by Houtgast and Steeneken.

There seems to be something special about the front ends of syllables. First, languages prefer CVs to VCs. Second, what children begin with is strings of CV-like pseudosyllables that emulate the syllable onsets of adult speech. Third there is perceptually significant information for the listener in the initial dynamics of the syllable. Let us add another phenomenon to this list: the *syllable beat*, or the syllable's P-center [16.118, 119].

In reading poetry, or in singing, we have a very strong sense that the syllables are spoken/sung in accordance with the rhythmic pattern of the meter. Native speakers agree more or less on how many syllables there are in a word or a phrase. In making such judgments, they seem to experience syllables as unitary events. Although it may take several hundred milliseconds to pronounce, subjectively a syllable appears to occur at a specific moment in time. It is this impression to which the phonetic term *syllable beat* refers and that has been studied experimentally in a sizable number of publications [16.120–126].

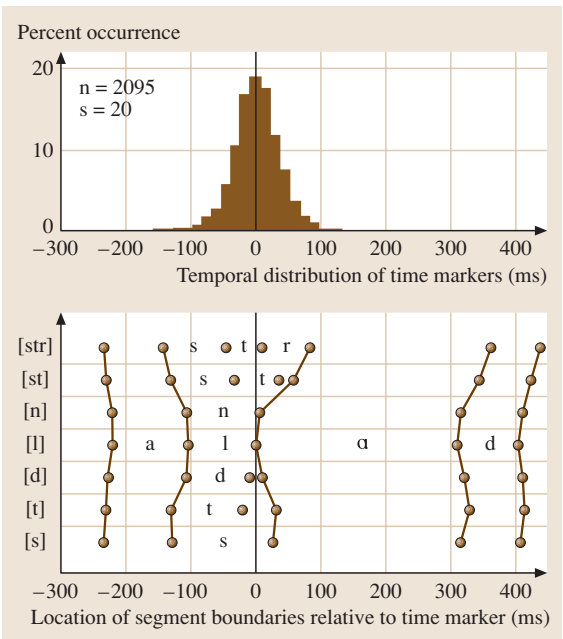


Fig. 16.37 In *Rapp* [16.100] three male Swedish subjects were asked to produce test words based on an [a_ˈɑːd] frame with [s, t, d, l, n, st, str] as intervocalic segment(s). These items were pronounced in synchrony with metronome beats presented over headphones. The figure shows the temporal distribution of the metronome beats (*upper panel*) in relation to the average time location of acoustic segment boundaries (*lower panel*)

Rapp [16.100] asked three native speakers of Swedish to produce random sets of test words built from [aC^{l}ɑːd] were the consonant C was selected from [s, t, d, l, n, st, str]. The instruction was to synchronize the stressed syllable with a metronome beat presented over earphones.

The results are summarized in Fig. 16.37. The *x*-axis represents distance in milliseconds from the point of reference, the metronome beat. The top diagram shows the total distribution of about 2000 time markers around the mean. The lower graph indicates the relative location of the major acoustic segment boundaries.

Several phonetic correlates have been proposed for the syllable beat: some acoustic/auditory, others articulatory. They all hover around the vowel onset, e.g., the amplitude envelope of a signal [16.127], rapid increase in energy in spectral bands [16.128, 129] or the onset of articulatory movement towards the vowel [16.130].

Rapp's data in Fig. 16.37 indicate that the mean beat time tends to fall close to the release or articulatory opening in [t, d, l, n] but that it significantly precedes the acoustic vowel onsets of [str-], [st-] and [s-]. However, when segment boundaries were arranged in relation to a fixed landmark on the F0 contour and vowel onsets were measured relative to that landmark, the range of vowel onsets was reduced. It is possible that the syllable beat may have its origin, not at the acoustic surface, nor at some kinematic level, but in a deeper motor control process that coordinates and imposes coherence on respiratory, phonatory and articulatory activity needed to produce a syllable.

Whatever the definitive explanation for the syllable's psychological *moment of occurrence* will be, the syllable beat provides a useful point of entry for attempts to understand how the control of rhythm and pitch works in speech and singing. Figure 16.38 compares spectrograms of the first few bars of *Over the rainbow*, spoken (left) and sung (right).

Vertical lines have been drawn at vowel onsets and points where articulators begin to move towards a more open configuration. The lines form an isochronous temporal pattern in the sung version which was performed at a regular rhythm. In the spoken example, they occur at intervals that seem more determined by the syllable's degree of prominence.

The subject reaches F0 targets at points near the beats (the vertical lines). From there target frequencies are maintained at stationary values until shortly before it is time to go to the next pitch. Thus the F0 curve resembles a step function with some smoothing applied to the steps.

On the other hand, the F0 contour for speech shows no such steady states. It makes few dramatic moves as it

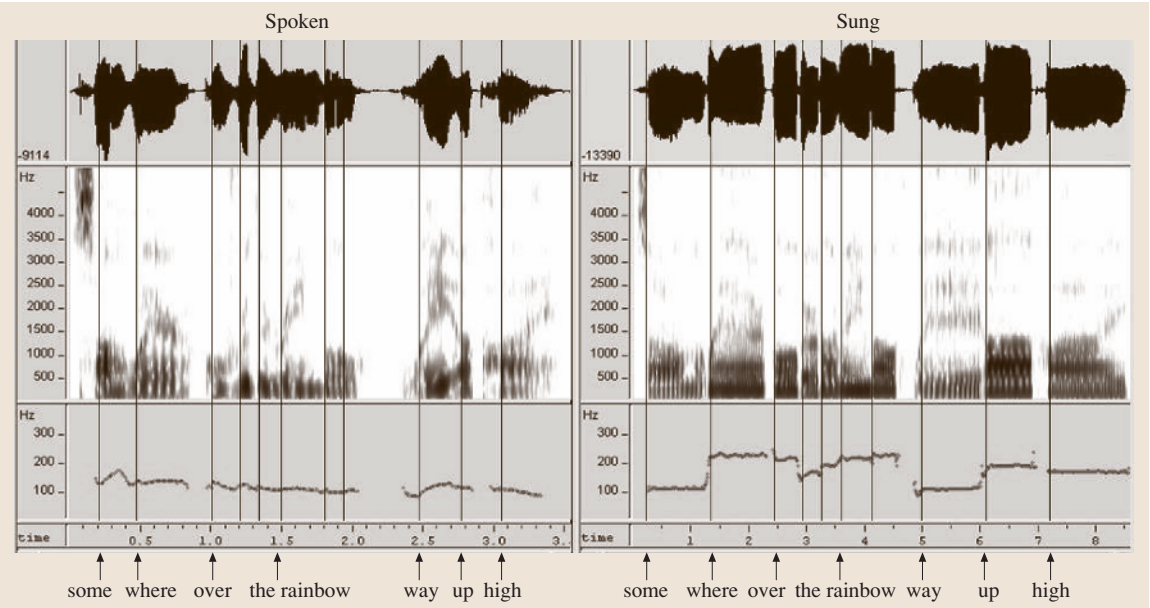


Fig. 16.38 Waveforms and spectrograms of the first few bars of Over the rainbow, spoken (*left*) and sung (*right*). Below: F0 traces in Hz. Vertical lines were drawn at time points corresponding to "vowel onsets" defined in terms of onset of voicing after a voiceless consonant, or the abrupt vocal tract opening following a consonant

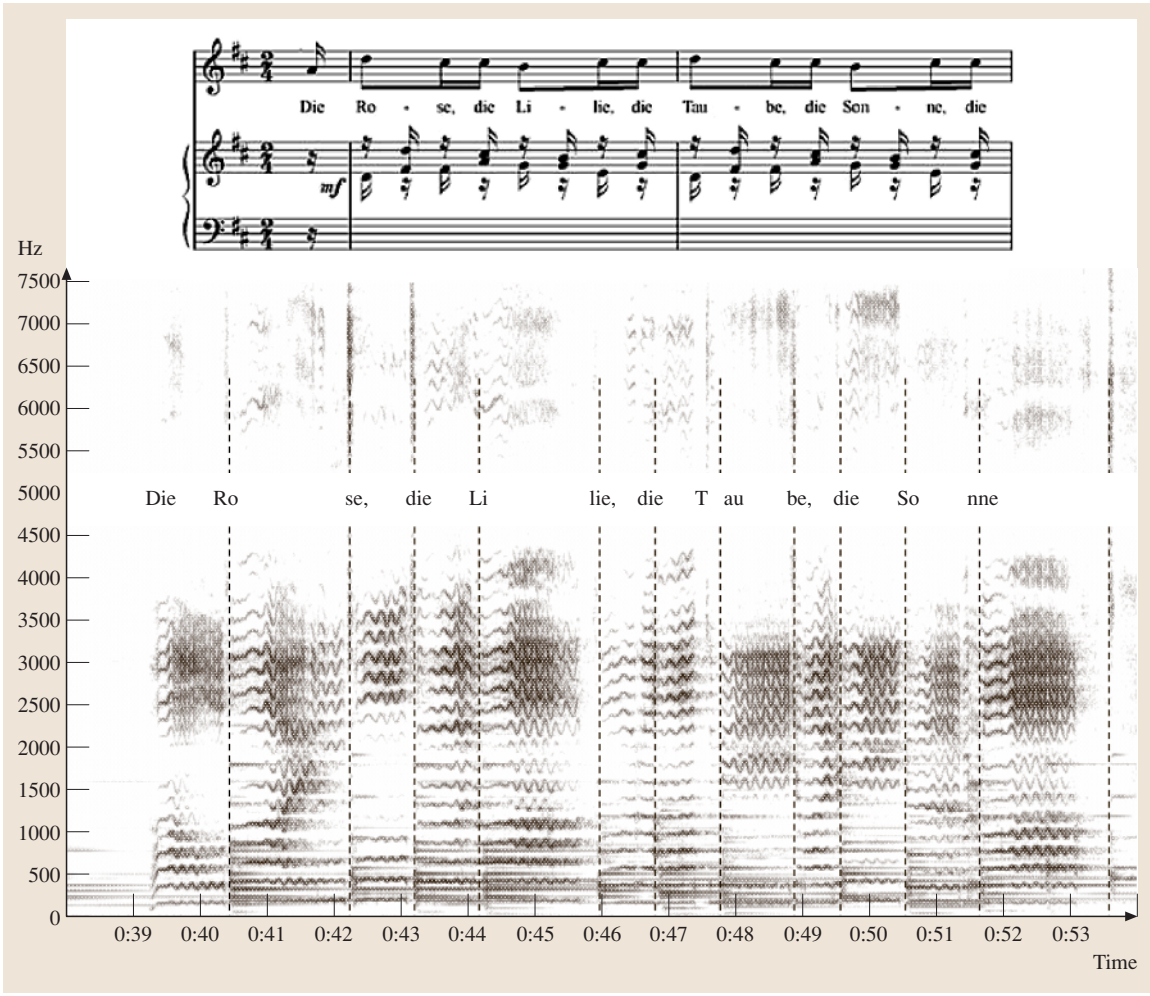


Fig. 16.39 Spectrogram of Dietrich Fischer-Dieskau's recording of song number 4 from Robert Schumann's Dichterliebe, op. 48. The *vertical bars* mark the onset of the tones of the piano accompaniment. Note the almost perfect synchrony between vowel onset and the piano. Note also the simultaneous appearance of high and low partials after the consonants also after the unvoiced /t/ in Taube

gradually drifts downward in frequency (the declination effect).

Figure 16.39 shows a typical example of classical singing, a spectrogram of a commercial recording of Dietrich Fischer-Dieskau's rendering of the song "*Die Rose, die Lilie*" from Robert Schumann's *Dichterliebe*, op. 48. The vertical dashed lines show the onsets of the piano accompaniment. The wavy patterns, often occurring somewhat after the vowel onset, reflect the vibrato. Apart from the vibrato undulation of the partials the gaps in the pattern of harmonics are quite apparent. At these points we see the effect of the more constricted articulations for the consonants. Note the rapid and crisply synchronous amplitude rise in all partials at the end of the consonant segments, also after unvoiced consonants, e.g. the /t/ in *Taube*. These rises are synchronized with the onset of the piano accompaniment tones, thus demonstrating that in singing a beat is marked by the vowel. The simultaneous appearing of low and high partials seems to belong to the characteristics of classical singing as opposed to speech, where the higher partials often arrive with a slight delay after unvoiced consonants. As mentioned before, such consonants are produced with abduction of the vocal folds. To generate

Amazon EX1014 Page 706

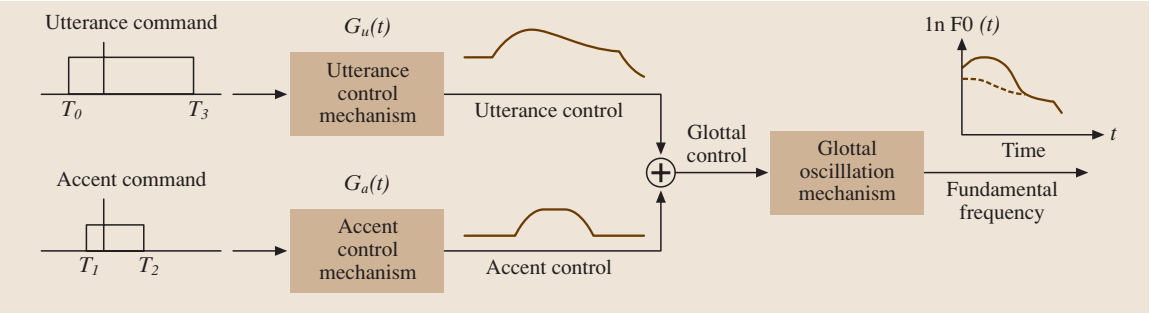


Fig. 16.40 A quantitative approach to synthesizing F0 contours by rule (After *Fujisaki* [16.131]). The framework shown here is an example of the so-called *superposition models* which generate F0 patterns by adding strongly damped responses to input step function target commands for phrase and accents

also high partials in this context, the vocal folds need to close the glottis at the very first vibratory cycle at the vowel onset. A potential benefit of this might be that this enhances the higher formants which are important to text intelligibility.

Several quantitative frameworks have been proposed for generating speech F0 contours by rule (for overview see Frid [16.132]). A subgroup of these has been called *superposition models* [16.131, 133]. A characteristic of such models is that they decompose the F0 curve into separate phrase and accent components. The accent commands are F0 step or impulse functions temporally linked to the syllables carrying stress. Accent pulses are superimposed on the phrase component. For a declarative sentence, the phrase component is a falling F0 contour produced from rectangular step function *hat patterns* [16.134]. The phrase and accent commands are passed through critically damped filters to convert their sum into the smoothly varying output F0 contour. An example of this approach is shown in Fig. 16.40 [16.131].

The last few figures suggest that the F0 patterns of singing and speech may be rather different if compared in terms of the raw F0 curves. However, the superposition models suggest that speech F0 contours have characteristics in part attributable to the passive response characteristics of the neuro-mechanical system that produces them, and in part due to active control signals. These control commands take the form of stepwise changes, some of short, others of longer duration. This representation is not unlike the sequence of F0 targets of a melody.

The implication of this analysis is that F0 is controlled in similar ways in speech and singing in the sense that both are based on sequences of underlying steady-state targets. On the other hand, a significant difference is that in singing high accuracy in attainment of acoustic target frequencies is required whereas in speech such demands are relaxed and smoothing is stronger.

16.6 Rhythm and Timing

A striking characteristic of a foreign language is its rhythm. Phoneticians distinguish between stress-timed and syllable-timed languages. English, Russian, Arabic and Thai are placed in the first group [16.135]. French, Spanish, Greek, Italian, Yoruba and Telugu are examples of the second.

Stress timing means that stressed syllables recur at approximately equal intervals. For instance, it is possible to say "*in ENGlish STRESses reCUR at EQual INtervals*", spacing the stressed syllables evenly in time and without sounding too unnatural. Stress increases syllable duration. In case several unstressed syllables occur in between stresses they are expected to undergo compensatory shortening. Syllable, or *machine-gun*, timing [16.136] implies that syllables recur at approximately equal intervals. In the first case it is the stresses that are isochronous, in the second the syllables.

To what extent are speech rhythms explicitly controlled by the speaker? To what extent are they fortuitous by-products of other factors? Does acquiring a new language involve learning new rhythmic target patterns?

The answer depends on how speech is defined. In the reading of poetry and nursery rhymes it is

Amazon EX1014 Page 707

clear that an external target, viz. the metric pattern, is a key determinant of how syllables are produced. Similarly, the groupings and the pausing in narrative prose, as read by a professional actor, can exhibit extrinsic rhythmic stylization compared with unscripted speech [16.137].

Dauer [16.138] points the way towards addressing such issues presenting statistics on the most frequently occurring syllable structures in English, Spanish and French. In Spanish and French, syllables were found to be predominantly open, i. e. ending with a vowel, whereas English showed a preference for closed syllables, i. e. ending with a consonant, especially in stressed syllables. Duration measurements made for English and Thai (stress-timed) and Greek, Spanish and Italian (syllable-timed) indicated that the duration of the interstress intervals grew at the same constant rate as a function of the number of syllables between the interstress intervals. In other words, no durational evidence was found to support the distinction between stress timing and syllable timing. How do we square that finding with the widely shared impression that some languages do indeed sound *stress-timed* and others *syllable-timed*?

Dauer [16.138, p. 55] concludes her study by stating that: " . . . *the rhythmic differences we feel to exist between languages such as English and Spanish are more a result of phonological, phonetic, lexical and syntactic facts about that language than any attempt on the part of the speaker to equalize interstress or intersyllable intervals*."

It is clear that speakers are certainly capable of imposing a rhythmic template in the serial read-out of syllables. But do they put such templates in play also when they speak spontaneously? According to *Dauer* and others [16.139, 140] rhythm is not normally an explicitly controlled variable. It is better seen as an emergent product of interaction among the choices languages make (and do not make) in building their syllables: e.g., from open versus closed syllable structures, heavy versus weak clusters, length distinctions and strong stressed/unstressed contrast. We thus learn that the distinction between syllable timing and stress timing may be a useful descriptive term but should primarily be applied to the phonetic output, more seldom to its input control.

Do speech rhythms carry over into music? In other words, would music composed by speakers of syllable-timed or stress-timed languages also be syllable timed or stress timed? The question has been addressed [16.141, 142] and some preliminary evidence has been reported.

There is more to speech timing than what happens inside the syllable. Processes at the word and phrase levels also influence the time intervals between syllables. Consider the English words *digest*, *insult* and *pervert*. As verbs their stress pattern can be described as a sequence of *weak–strong* syllables. As nouns the order is reversed: *strong–weak*. The syllables are the same but changing the word's stress contour (the lexical stress) affects their timing.

The *word length effect* has been reported for several languages [16.58, 143]. It refers to the tendency for the stressed vowel of the word to shorten as more syllables are appended, cf. English *speed*, *speedy*, *speedily*. In Lehiste's formulation: "It appears that in some languages the word as a whole has a certain duration that tends to remain relatively constant, and if the word contains a greater number of segmental sounds, the duration of the segmental sounds decreases as their number in the word increases."

At the phrase level we find that rhythmic patterns can be used to signal differences in syntactic structure. Compare:

1. *The 2000-year-old skeletons*,
2. *The two 1000-year-old skeletons*.

The phrases contain syllables with identical phonetic content but are clearly timed differently. The difference is linked to that fact that in (1) *2000-year-old* forms a single unit, whereas in (2) *two 1000-year-old* consists of two constituents.

A further example is:

3. *Johan greeted the girl with the flowers*.

Hearing this utterance in a specific context a listener might not find it ambiguous. But it has two interpretations: (a) either Johan greeted the girl who was carrying flowers, or (b) he used the flowers to greet her. If spoken clearly to disambiguate, the speaker is likely to provide temporal cues in the form of shortening the syllables within each syntactic group to signal the coherence of the constituent syllables (cf. the word length effect above) and to introduce a short pause between the two groups. There would be a lengthening of the last syllable before the pause and of the utterance-final syllable.

Version (a) can be represented as

4. [*Johan greeted*] # [*the girl with the flowers*].

In (b) the grouping is

5. [*Johan greeted the girl*] # [*with the flowers*].

The # symbol indicates the possibility of a short juncture or pause and a lengthening of the segments preceding the pause. This boundary cue is known as *pre-pausal lengthening*, or more generally as *final lengthening* [16.143, 144], a process that has been observed for a great many languages. It is not known whether this process is a language universal. It is fair to say that is typologically widespread.

Curiously it resembles a phenomenon found in poetry, folk melodies and nursery tunes, called *catalexis*. It consists in omitting the last syllable(s) in a line or other metrical unit of poetry. Instead of four mechanically repeated trochaic feet as in:

|– ⏑ |– ⏑ |– ⏑ |– ⏑ |
|– ⏑ |– ⏑ |– ⏑ |– ⏑ |

we typically find catalectic lines with durationally implied but deleted final syllables as in:

|– ⏑ |– ⏑ |– ⏑ |– |

Old McDonald had a farm

|– ⏑ |– ⏑ |– | |

ee-i ee-i oh

Final lengthening is an essential feature of speech prosody. Speech synthesized without it sounds both highly unnatural and is harder to perceive. In music performance frameworks, instrumental as well as sung, final lengthening serves the purpose of grouping and constituent marking [16.145, 146].

16.7 Prosody and Speech Dynamics

Degree of prominence is an important determinant of segment and syllable duration in English. Figure 16.41 shows spectrograms of the word '*squeal*' spoken with four degrees of stress in sentences read in response to a list of questions (source: [16.147]). The idea behind this method was to elicit tokens having emphatic, strong (as in focus position), neutral and weak (unfocused) stress on the test syllable. The lower row compares the strong and the emphatic pronunciations (left and right respectively). The top row presents syllables with weaker degrees of stress.

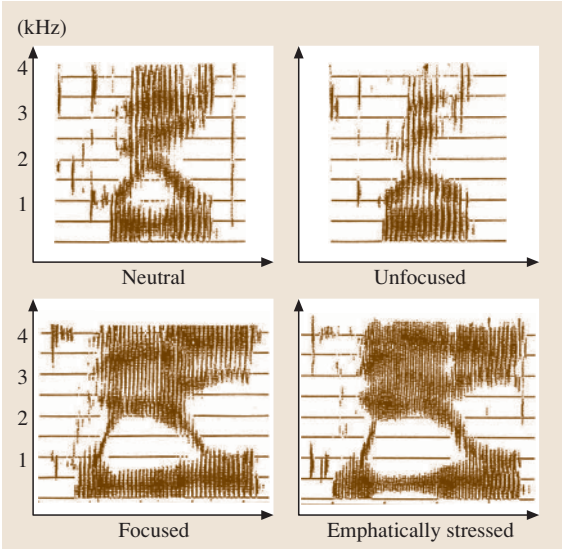


Fig. 16.41 Spectrograms of the word 'squeal' spoken with four degrees of stress in sentences read in response to a list of questions. (After *Brownlee* [16.147])

The representations demonstrate that the differences in stress have a marked effect on the word's spectrographic pattern. Greater prominence makes it longer. Formants, notably F2, show more extensive and more rapid transitions.

A similar experimental protocol was used in two studies of the dynamics of vowel production [16.147, 148]. Both report formant data on the English front vowels [i], [ɪ], [ɛ] and [eɪ] occurring in syllables selected to maximize and minimize contextual assimilation to the place of articulation of the surrounding consonants. To obtain a maximally assimilatory frame, words containing the sequence [w_ɬ] were chosen, e.g., as in *wheel*, *will*, *well* and *wail*. (The [w] is a labio-velar and English [ɬ] is velarized. Both are thus made with the tongue in a retracted position). The minimally assimilatory syllable was [h_d].

Moon's primary concern was *speaking style* (clear versus casual speech). Brownlee investigated changes due to *stress variations*. In both studies measurements were made of vowel duration and extent of formant transitions. Vowel segment boundaries were defined in terms of fixed transition onset and endpoint values for [w] and [ɬ] respectively. The [h_d] frame served as a *null context* reference. The [w_ɬ] environment produced formant transitions large enough to provide a sensitive and robust index of articulatory movement (basically the front–back movement of the tongue).

Amazon EX1014 Page 709

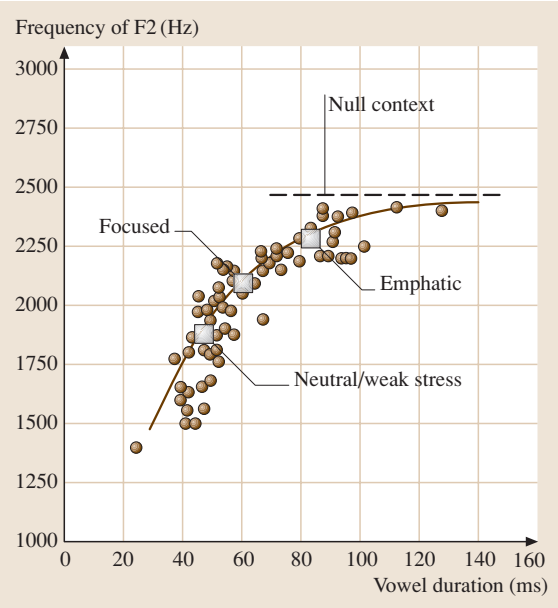


Fig. 16.42 Measurements of F2 as function of vowel duration from tokens of [i] in the word 'squeal' spoken by a male subject under four conditions of stress: emphatically stressed, focused, neutral stress and weak stress (out of focus). *Filled circles* pertain to the individual measurements from all the four stress conditions. The *dashed horizontal line*, is the average F2 value of citation form productions of [h_d]. The averages for the four stress conditions are indicated by the *open squares*. As duration increases – which is approximately correlated with increasing stress – the *square symbols* are seen to move up along the smooth curve implying decreasing undershoot. (After *Brownlee* [16.147])

Figure 16.42 presents measurements of F2 and vowel duration from tokens of [i] in the word '*squeal*' spoken by a male subject under four conditions of stress: emphatically stressed, focused, neutral stress and weak stress (out of focus). The subject was helped to produce the appropriate stress in the right place by reading a question before each test sentence.

Filled circles pertain to individual measurements pooled for all stress conditions. The points form a coherent cluster which is fairly well captured by an exponential curve. The asymptote is the dashed horizontal line, i. e., the average F2 value of the [h_d] data for [i]. The averages for the stress conditions are indicated by the unfilled squares. They are seen to move up along the curve with increasing stress (equivalently, vowel duration).

To understand the articulatory processes underlying the dynamics of vowel production it is useful to adopt a biomechanical perspective. In the [w_ɬ] test words the tongue body starts from a posterior position in [w] and moves towards a vowel target located in the front region of the vocal tract, e.g., [i], [ɪ], [ɛ] or [e]. Then it returns to a back configuration for the dark [ɬ]. At short vowel durations there is not enough time for the vowel gesture to be completed. As a result, the F2 movement misses the reference target by several hundred Hz. Note that in unstressed tokens the approach to target is off by almost an octave. As the syllable is spoken with more stress, and the vowel accordingly gets longer, the F2 transition falls short to a lesser degree. What is illustrated here is the phenomenon known as *formant undershoot* [16.149, 150].

Formant undershoot has received a fair amount of attention in the literature [16.151–155]. It is generally seen as an expression of the sluggish response characteristics of the speech production system. The term *sluggish* here describes both neural delays and mechanical time constants. The response to the neural commands for a given movement is not instantaneous. It takes time for neural impulses to be transformed into muscular contractions and it takes time for the tongue, the jaw and other articulatory structures to respond mechanically to the forces generated by those contractions. In other words, several stages of filtering take place between control signals and the unfolding of the movements. It is this filtering that makes the articulators sluggish. When commands for successive phonemes arrive faster than the articulatory systems are able to respond, the output is a set of incomplete movements. There is undershoot and failure to reach spatial, and thus also acoustic, targets.

However, biomechanics tells us that, in principle, it should be possible to compensate for system characteristics by applying greater force to the articulators thereby speeding up movements and improving target attainment [16.156]. Such considerations make us expect that, in speech movement dynamics, a given trajectory is shaped by

1. the distance between successive articulatory goals (the extent of movement);
2. articulatory effort (input force); and
3. duration (the time available to execute the movement).

The data of Brownlee are compatible with such a mechanism in that the stress-dependent undershoot observed is likely to be a joint consequence of: (1) the large distances between back and front targets; (2) the stress differences corresponding to variations in articulatory effort; and (3) durational limitations.

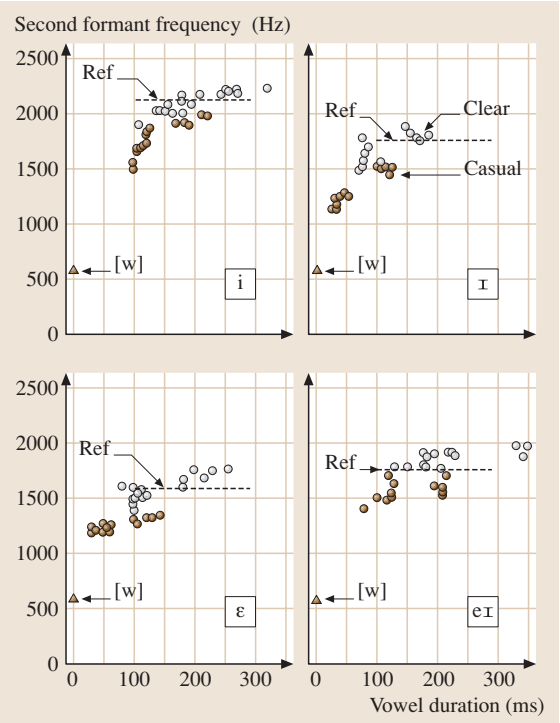


Fig. 16.43 Measurements F2 as function of vowel duration for five male subjects' casual citation-form style pronunciation of the words *wheel*, *wheeling*, *Wheelingham*, *will*, *willing*, *willingly* and for their pronunciation of the same words when asked to say them clearly in an overarticulated way. Casual style is represented by the *solid circles*, clear speech by *open circles*. The frequency of F2 in [w] is indicated by the arrows at 600 Hz. The mean F2 for the [h_d] data is entered as dashed horizontal lines. (After [16.157], Fig. 4)

The three factors – movement extent, effort and duration – are clearly demonstrated in Moon's research on clear and casual speech. As mentioned, Moon's test syllables were the same as Brownlee's. He varied speaking style while keeping stress constant. This was achieved by taking advantage of the word-length effect.The segment strings [wiɬ-], [wɪɬ-], [wɛɬ-] and [weɬ-] were used as the first main-stressed syllable in mono-, bi- and trisyllabic words to produce items such as *wheel*, *wheeling*, *Wheelingham, will*, *willing*, *willingly*, etc.

In the first part of his experiment, five male subjects were asked to produce these words in casual citation-form style. In the second section the instruction was to say them clearly in an *overarticulated* way. For the citation-form pronunciation the results replicated previously observed facts in that: (a) all subjects exhibited duration-dependent formant undershoot, and (b) the magnitude of this effect varied in proportion to the extent of the [w]–[vowel] transition. In the clear style, undershoot effects were reduced. The mechanism by which this was achieved varied somewhat from subject to subject. It involved combinations of increased vowel duration and more rapid and more extensive formant transitions. Figure 16.43 shows the response of one of the subjects to the two tasks [16.157].

Casual style is represented by the solid circles, clear speech by open circles. The frequency of F2 in [w] is indicated by the arrows at 600 Hz. The mean F2 for the [h_d] data is entered as dashed horizontal lines.

The solid points show duration-dependent undershoot effects in relation to the reference values for the [h_d] environment. The open circles, on the other hand, overshoot those values suggesting that this talker used more extreme F2 targets for the clear condition. The center of gravity of the open circles is shifted to the right for all four vowels showing that clear forms were longer than citation forms. Accordingly, this is a subject who used all three methods to decrease undershoot: clear pronunciations exhibited consistently more extreme targets (F2 higher), longer durations and more rapid formant movements.

These results demonstrate that duration and context are not the only determinants of vowel reduction since, for any given duration, the talker is free to vary the degree of undershoot by choosing to articulate more forcefully (as in overarticulated *hyperspeech*) or in a more relaxed manner (as in casual hypospeech). Taken together, the studies by Moon and Brownlee suggest that the dynamics of articulatory movement, and thus of formant transitions, are significantly constrained by three factors: extent of movement, articulatory effort and duration.

16.8 Control of Sound in Speech and Singing

When we perform an action – walk, run, or reach for and manipulate objects – our motor systems are faced with the fact that the contexts under which movements are made are never exactly the same. They change significantly from one situation to the next. Nonetheless, motor systems adapt effortlessly to the continually changing

conditions presumably because, during evolution, they were shaped by the need to cope with unforeseen events and obstacles [16.158]. Their default mode of operation is compensatory [16.159].

Handwriting provides a good illustration of this ability. A familiar fact is that it does not matter if something is written on the blackboard or on a piece of paper. The characteristic features of someone's handwriting are nevertheless easily recognized. Different sets of muscles are recruited and the size of the letters is different but their shapes remain basically similar. What this observation tells us is that movements are not specified in terms of fixed set of muscles and constant contraction patterns. They are recruited in functionally defined groups, *coordinative structures* [16.160]. They are planned and executed in an external coordinate space, in other words in relation to the 3-D world in which they occur so as to attain goals defined outside the motor system itself. The literature on motor mechanisms teaches us that voluntary movements are prospectively organized or future-oriented.

Speech and singing provide numerous examples of this output-oriented mode of motor control [16.161–163]. Earlier in the chapter we pointed out that, in upright position, the diaphragm and adjacent structures are influenced by gravity and tend to be pulled down, thereby causing the volume of the thoracic cavity to increase. In this position, gravity contributes to the inspiratory forces. By contrast, in the supine position, the diaphragm tends to get pushed up into the rib cage, which promotes expiration [16.1, p. 40].

Sundberg et al. [16.164] investigated the effect of upright and supine positions in two baritone singers using synchronous records of esophageal and gastric pressure, EMG from inspiratory and expiratory muscles, lung volume and sound. Reorganization of respiratory activity was found and was interpreted as compensation for the different mechanical situations arising from the upright and supine conditions.

This finding is closely related to what we know about breathing during speech. As mentioned above (Fig. 16.6), the Ps tends to stay fairly constant for any given vocal effort. It is known that this result is achieved by tuning the balance between inspiratory and expiratory muscles. When the lungs are expanded so that the effect of elastic recoil creates a significant expiration force, inspiratory muscles predominate to put a brake on that force. For reduced lung volumes the situation is the opposite. The effect of elastic recoil is rather to increase lung volume. Accordingly, the muscle recruitment needed to maintain Ps is expected to be primarily expiratory. That is indeed what the data show [16.8, 66, 165].

The bite-block paradigm offers another speech example. In one set of experiments [16.166] subjects were instructed to pronounce syllables consisting only of a long vowel under two conditions: first normally, then with a bite block (BB) between the teeth. The subjects all non-phoneticians were told to try to sound as normal as possible despite the BB. No practice was allowed. The purpose of the BB was to create an abnormally large jaw opening for close vowels such as [i] and [u]. It was argued that, if no tongue compensation occurred, this large opening would drastically change the area function of the close vowels and disrupt their formant patterns. In other words, the question investigated was whether the subjects were able to sound normal despite the BB.

Acoustic recordings were made and formant pattern data were collected for comparisons between conditions, vowels and subjects. The analyses demonstrated clearly that subjects were indeed able to produce normal sounding vowels in spite of the BB. At the moment of the first glottal pulse formant patters were well within the normal ranges of variation.

In a follow-up X-ray investigation [16.167] it was found that compensatory productions of [i] and [u] were made with super-palatal and super-velar tongue shapes. In other words, tongue bodies were raised high above normal positions so as to approximate the normal cross-sectional area functions for the test vowels.

Speech and singing share a lot of features but significant differences are brought to light when we consider what the goals of the two behaviors are.

It is more than 50 years since the sound spectrograph became commercially available [16.91]. In this time we have learned from visible speech displays and other records that the relation between phonetic units and the acoustic signal is extremely complex [16.168]. The lesson taught is that invariant correlates of linguistic categories are not readily apparent in the speech wave [16.169]. In the preceding discussion we touched on some of the sources of this variability: coarticulation, reduction and elaboration, prosodic modulation, stylistic, situational and speaker-specific characteristics. From this vantage point it is astonishing to note that, even under noisy conditions, speech communication is a reliable and robust process. How is this remarkable fact to be accounted for?

In response to this problem a number of ideas and explanatory frameworks have been proposed. For instance, it has been suggested that acoustic invariants are rela-

tional rather than absolute (à la tone intervals defined as frequency ratios).

Speech is rather a set of movements made audible than a set of sounds produced by movements [16.14, p. 33].

In keeping with this classical statement, many investigators have argued that speech entities are to be found at some upstream speech production level and should be defined as *gestures* [16.170–173].

At the opposite end, we find researchers (e.g., Perkell [16.163]) who endorse Roman Jakobson's view which, in the search for units gives primacy to the perceptual representation of speech sounds. To Jakobson the stages of the speech chain form an "... *operational hierarchy of levels of decreasing pertinence: perceptual, aural, acoustical and articulatory (the latter carrying no direct information to the receiver)*." [16.174]

These viewpoints appear to conflict and do indeed divide the field into those who see speech as a motoric code (the *gesturalist* camp) and those who maintain that it is primarily shaped by perceptual processes (the *listener-oriented* school).

There is a great deal of experimental evidence for both sides, suggesting that this dilemma is not an either–or issue but that both parties offer valuable complementary perspectives.

A different approach is taken by the H & H (Hyper and Hypo) theory [16.175, 176]. This account is developed from the following key observations about speaker–listener interaction:

1. Speech perception is always a product of signal information and listener knowledge;
2. Speech production is adaptively organized.

Here is an experiment that illustrates the first claim about perceptual processing. Two groups of subjects listen to a sequence of two phrases: a question followed by an answer. The subject groups hear different questions but a single physically identical reply. The subjects' task is to say how many words the reply contains.

The point made here is that Group 1 subjects hear [lɛsn̩ faɪv] as "*less than five*". Those in Group 2 interpret it as "*lesson five*". The first group's response is "*three words*", and the answer of the second is "*two words*". This is *despite the fact* that physically the [lɛsn̩ faɪv] stimulus is exactly the same. The syllabic [n̩] signals the word *than* in one case and the syllable *–on* in the other. Looking for the invariant correlates of the initial consonant *than* is doomed to failure because of the severe degree of reduction.

To proponents of H & H theory this is not an isolated case. This is the way that perception in general works. The speech percepts can never be raw records of the signal because listener knowledge will inevitably interact with the stimulus and will contribute to shaping the percept.

Furthermore, H & H theory highlights the fact that spoken messages show a non-uniform distribution of information in that predictability varies from situation to situation, from word to word and from syllable to syllable. Compare (a) and (b) below. What word is likely to be represented by the gap?

1. "*The next word is* __."
2. "*A bird in the hand is worth two in the* __."

Any word can be expected in (1) whereas in (2) the predictability of the word "*bush*" is high.

H & H theory assumes that, while learning and using their native language, speakers develop a sense of this informational dynamics. Introducing the abstraction of an *ideal speaker*, it proposes that the talker estimates the running contribution that signal-complementary information (listener knowledge) will make during the course of the utterance and then tunes his articulatory performance to the presumed short-term listener needs. Statistically, this type of behavior has the long-term consequence of distributing phonetic output forms along a continuum with clear and elaborated forms (hyperspeech) at one end and casual and reduced pronunciations (hypospeech) at the other.

The implication for the invariance issue is that the task of the speaker is not to encode linguistic units as physical invariants, but to make sure that the signal attributes (of phonemes, syllables, words and phrases) carry discriminative power *sufficient* for successful lexical access. To make the same point in slightly different terms, the task of the signal is not to embody phonetic patterns of constancy but to provide *missing information*.

It is interesting to put this account of speech processes next to what we know about singing. Experimental observations indicate that singing in tune is not simply about invariant F0 targets. F0 is affected by the singer's expressive performance [16.177] and by the tonal context in which a given note is embedded [16.178]. Certain deviations from nominal target frequencies are not unlike the coarticulation and undershoot effects ubiquitously present in speech. Accordingly, with respect to frequency control, speaking and singing are qualitatively similar. However, quantitatively, they differ drastically. Recall that in describing the dynamics of vowel reduction we noted that formant fre-

Amazon EX1014 Page 713

quencies can be displaced by as much as 50% from target values. Clearly, a musician or singer with a comparable under-/overshoot record would be recommended to embark on an alternative career, the margins of perceptual tolerance being much narrower for singing.

What accounts for this discrepancy in target attainment? Our short answer is that singing or playing *out of tune* is a bad thing. Where does this taboo come from? From consonance and harmony constraints. In simplified terms, an arbitrary sample of tonal music can be analyzed into a sequence of chords. Its melodic line is a rhythmic and tonal elaboration of this harmonic structure. Statistically, long prominent melody tones tend to attract chord notes. Notes of less prominence typically *interpolate* along scales in smaller intervals between the metrically heavier chord notes. The notion of *consonance* goes a long way towards explaining why singing or playing out of tune is tabooed in music: hitting the right pitch is required by the consonance and harmony constraints expected by the listener and historically presumably linked with a combination of the polyphony in our Western tradition of music composition and the fact that most of our music instruments produce tones with harmonic spectra. This implies that intervals departing too much from just intonation will generate beats between simultaneously sounding and only nearly coinciding partials, particularly if the fundamental frequencies are constant, lacking vibrato and flutter.

16.9 The Expressive Power of the Human Voice

The human voice is an extremely expressive instrument both when used for speech and for singing. By means of subtle variations of timing and pitch contour speakers and singers add a substantial amount of expressiveness to the linguistic or musical content and we are quite skilled in deciphering this information. Indeed a good deal of vocal artistry seems to lie in the artist's skill in making nothing but such changes of pitch, timbre, loudness and timing that a listener can perceive as carrying some meaning.

We perceive the extra-linguistic or expressive information in speech and singing in various shapes. For example, we can interpret certain combinations of acoustic characteristics in speech in terms of a smile or a particular forming of the lips on the face of the speaker [16.179]. For example *Fónagy* [16.180] found that listeners were able to replicate rather accurately the facial expression of speakers only by listening to their voices.

The emotive transforms in speech and singing seem partly similar and sometimes identical [16.181, 182]. Final lengthening, mentioned above, uses the same code for marking the end of a structural element, such as a sentence in speech or a phrase in sung and played music performance. Emphasis by delayed arrival is another example, i. e., delaying an emphasized stressed syllable or note by lengthening the unstressed syllable/note preceding it [16.183].

The expressive potential of the human voice is indeed enormous, and would transpire from the ubiquitous use of the voice for the purpose of communication. Correct interpretation of the extra-linguistic content of a spoken utterance is certainly important in our daily life, so we are skilled in deciphering vocal signals also along those dimensions. The importance of correct encoding of the extra-linguistic implies that speakers acquire a great skill in this respect. This skill would be the basic requirement for vocal art, in singing as well as in acting.

References

16.1 T.J. Hixon: *Respiratory Function in Speech and Song* (Singular, San Diego 1991) pp. 1–54

16.2 A. L. Winkworth, P. J. Davis, E. Ellis, R. D. Adams: Variability and consistency in speech breathing during reading: Lung volumes, speech intensity, and linguistic factors, JSHR **37**, 535–556 (1994)

16.3 M. Thomasson: *From Air to Aria*, Ph.D. Thesis (Music Acoustics, KTH 2003)

16.4 B. Conrad, P. Schönle: Speech and respiration, Arch. Psychiat. Nervenkr. **226**, 251–268 (1979)

16.5 M.H. Draper, P. Ladefoged, D. Whitteridge: Respiratory muscles in speech, J. Speech Hear. Disord. **2**, 16–27 (1959)

16.6 R. Netsell: Speech physiology. In: *Normal aspects of speech, hearing, and language*, ed. by P.D. Minifie, T.J. Hixon, P. Hixon, P. Williams (Prentice-Hall, Englewood Cliffs 1973) pp. 211–234

16.7 P. Ladefoged, M.H. Draper, D. Whitteridge: Syllables and stress, Misc. Phonetica **3**, 1–14 (1958)

16.8 P. Ladefoged: Speculations on the control of speech. In: *A Figure of Speech: A Festschrift for John Laver*, ed. by W.J. Hardcastle, J. Mackenzie Beck (Lawrence Erlbaum, Mahwah 2005) pp. 3–21

16.9 T.J. Hixon, G. Weismer: Perspectives on the Edinburgh study of speech breathing, J. Speech Hear. Disord. **38**, 42–60 (1995)

16.10 S. Nooteboom: The prosody of speech: melody and rhythm. In: *The Handbook of Phonetic Sciences*, ed. by W.J. Hardcastle, J. Laver (Blackwell, Oxford 1997) pp. 640–673

16.11 M. Rothenberg: *The breath-stream dynamics of simple-released-plosive production Bibliotheca Phonetica 6* (Karger, Basel 1968)

16.12 D.H. Klatt, K.N. Stevens, J. Mead: Studies of articulatory activity and airflow during speech, Ann. NY Acad. Sci. **155**, 42–55 (1968)

16.13 J.J. Ohala: Respiratory activity in speech. In: *Speech production and speech modeling*, ed. by W.J. Hardcastle, A. Marchal (Dordrecht, Kluwer 1990) pp. 23–53

16.14 R.H. Stetson: *Motor Phonetics: A Study of Movements in Action* (North Holland, Amsterdam 1951)

16.15 P. Ladefoged: Linguistic aspects of respiratory phenomena, Ann. NY Acad. Sci. **155**, 141–151 (1968)

16.16 L.H. Kunze: Evaluation of methods of estimating sub-glottal air pressure muscles, J. Speech Hear. Disord. **7**, 151–164 (1964)

16.17 R. Leanderson, J. Sundberg, C. von Euler: Role of the diaphragmatic activity during singing: a study of transdiaphragmatic pressures, J. Appl. Physiol. **62**, 259–270 (1987)

16.18 J. Sundberg, N. Elliot, P. Gramming, L. Nord: Short-term variation of subglottal pressure for expressive purposes in singing and stage speech. A preliminary investigation, J. Voice **7**, 227–234 (1993)

16.19 J. Sundberg: Synthesis of singing, in Musica e Technologia: Industria e Cultura per lo Sviluppo del Mezzagiorno. In: *Proceedings of a symposium in Venice*, ed. by R. Favaro (Unicopli, Milan 1987) pp. 145–162

16.20 J. Sundberg: Synthesis of singing by rule. In: *Current Directions in Computer Music Research, System Development Foundation Benchmark Series*, ed. by M. Mathews, J. Pierce (MIT, Cambridge 1989), 45–55 & 401–403

16.21 J. Molinder: *Akustiska och perceptuella skillnader mellan röstfacken lyrisk och dramatisk sopran, unpublished thesis work* (Lund Univ. Hospital, Dept of Logopedics, Lund 1997)

16.22 T. Baer: Reflex activation of laryngeal muscles by sudden induced subglottal pressure changes, J. Acoust. Soc. Am. **65**, 1271–1275 (1979)

16.23 T. Cleveland, J. Sundberg: Acoustic analyses of three male voices of different quality. In: *SMAC 83. Proceedings of the Stockholm Internat Music Acoustics Conf*, Vol. 1, ed. by A. Askenfelt, S. Felicetti, E. Jansson, J. Sundberg (Roy. Sw. Acad. Music, Stockholm 1985) pp. 143–156, No. 46:1

16.24 J. Sundberg, C. Johansson, H. Willbrand, C. Ytterbergh: From sagittal distance to area, Phonetica **44**, 76–90 (1987)

16.25 I.R. Titze: Phonation threshold pressure: A missing link in glottal aerodynamics, J. Acoust. Soc. Am. **91**, 2926–2935 (1992)

16.26 I.R. Titze: *Principles of Voice Production* (Prentice-Hall, Englewood Cliffs 1994)

16.27 G. Fant: *Acoustic theory of speech production* (Mouton, The Hague 1960)

16.28 K.N. Stevens: *Acoustic Phonetics* (MIT, Cambridge 1998)

16.29 M. Hirano: *Clinical Examination of Voice* (Springer, New York 1981)

16.30 M. Rothenberg: A new inversefiltering technique for deriving the glottal air flow waveform during voicing, J. Acoust. Soc. Am. **53**, 1632–1645 (1973)

16.31 G. Fant: The voice source – Acoustic modeling. In: *STL/Quart. Prog. Status Rep. 4* (Royal Inst. of Technology, Stockholm 1982) pp. 28–48

16.32 C. Gobl: *The voice source in speech communication production and perception experiments involving inverse filtering and synthesis. D.Sc. thesis* (Royal Inst. of Technology (KTH), Stockholm 2003)

16.33 G. Fant, J. Liljencrants, Q. Lin: A four-parameter model of glottal flow. In: *STL/Quart. Prog. Status Rep. 4, Speech, Music and Hearing* (Royal Inst. of Technology, Stockholm 1985) pp. 1–13

16.34 D.H. Klatt, L.C. Klatt: Analysis, synthesis and perception of voice quality variations among female and male talkers, J. Acoust. Soc. Am. **87**(2), 820–857 (1990)

16.35 M. Ljungqvist, H. Fujisaki: A comparative study of glottal waveform models. In: *Technical Report of the Institute of Electronics and Communications Engineers*, Vol. EA85-58 (Institute of Electronics and Communications Engineers, Tokyo 1985) pp. 23–29

16.36 A.E. Rosenberg: Effect of glottal pulse shape on the quality of natural vowels, J. Acoust. Soc. Am. **49**, 583–598 (1971)

16.37 M. Rothenberg, R. Carlson, B. Granström, J. Lindqvist-Gauffin: A three- parameter voice source for speech synthesis. In: *Proceedings of the Speech Communication Seminar 2*, ed. by G. Fant (Almqvist & Wiksell, Stockholm 1975) pp. 235–243

16.38 K. Ishizaka, J.L. Flanagan: Synthesis of voiced sounds from a two-mass model of the vocal cords, The Bell Syst. Tech. J. **52**, 1233–1268 (1972)

16.39 Liljencrants: Chapter A translating and rotating mass model of the vocal folds. In: *STL/Quart. Prog. Status Rep. 1, Speech, Music and Hearing* (Royal Inst. of Technology, Stockholm 1991) pp. 1–18

16.40 A. Ní Chasaide, C. Gobl: Voice source variation. In: *The Handbook of Phonetic Sciences*, ed. by

Amazon EX1014 Page 715

W.J. Hardcastle, J. Laver (Blackwell, Oxford 1997) pp. 427–462

16.41 E.B. Holmberg, R.E. Hillman, J.S. Perkell: Glottal air flow and pressure measurements for loudness variation by male and female speakers, J. Acoust. Soc. Am. **84**, 511–529 (1988)

16.42 J.S. Perkell, R.E. Hillman, E.B. Holmberg: Group differences in measures of voice production and revised values of maximum airflow declination rate, J. Acoust. Soc. Am. **96**, 695–698 (1994)

16.43 J. Gauffin, J. Sundberg: Spectral correlates of glottal voice source waveform characteristics, J. Speech Hear. Res. **32**, 556–565 (1989)

16.44 J. Svec, H. Schutte, D. Miller: On pitch jumps between chest and falsetto registers in voice: Data on living and excised human larynges, J. Acoust. Soc. Am. **106**, 1523–1531 (1999)

16.45 J. Sundberg, M. Andersson, C. Hultqvist: Effects of subglottal pressure variation on professional baritone singers voice sources, J. Acoust. Soc. Am. **105**, 1965–1971 (1999)

16.46 J. Sundberg, E. Fahlstedt, A. Morell: Effects on the glottal voice source of vocal loudness variation in untrained female and male subjects, J. Acoust. Soc. Am. **117**, 879–885 (2005)

16.47 P. Sjölander, J. Sundberg: Spectrum effects of subglottal pressure variation in professional baritone singers, J. Acoust. Soc. Am. **115**, 1270–1273 (2004)

16.48 P. Branderud, H. Lundberg, J. Lander, H. Djamshidpey, I. Wäneland, D. Krull, B. Lindblom: *X-ray analyses of speech: Methodological aspects*, *Proc. of 11th Swedish Phonetics Conference* (Stockholm Univ., Stockholm 1996) pp. 168–171

16.49 B. Lindblom: A numerical model of coarticulation based on a Principal Components analysis of tongue shapes. In: *Proc. 15th Int. Congress of the Phonetic Sciences*, ed. by D. Recasens, M. Josep Solé, J. Romero (Universitat Autònoma de Barcelona, Barcelona 2003), CD-ROM

16.50 G.E. Peterson, H. Barney: Control methods used in a study of the vowels, J. Acoust. Soc. Am. **24**, 175–184 (1952)

16.51 Hillenbrand et al.: Acoustic characteristics of American English vowels, J. Acoust. Soc. Am. **97**(5), 3099–3111 (1995)

16.52 G. Fant: Analysis and synthesis of speech processes. In: *Manual of Phonetics*, ed. by B. Malmberg (North-Holland, Amsterdam 1968) pp. 173–277

16.53 G. Fant: Formant bandwidth data. In: *STL/Quart. Prog. Status Rep. 7* (Royal Inst. of Technology, Stockholm 1962) pp. 1–3

16.54 G. Fant: Vocal tract wall effects, losses, and resonance bandwidths. In: *STL/Quart. Prog. Status Rep. 2–3* (Royal Inst. of Technology, Stockholm 1972) pp. 173–277

16.55 A.S. House, K.N. Stevens: Estimation of formant bandwidths from measurements of transient response of the vocal tract, J. Speech Hear. Disord. **1**, 309–315 (1958)

16.56 O. Fujimura, J. Lindqvist: Sweep-tone measurements of vocal-tract characteristics, J. Acoust. Soc. Am. **49**, 541–558 (1971)

16.57 I. Lehiste, G.E. Peterson: Vowel amplitude and phonemic stress in American English, J. Acoust. Soc. Am. **3**, 428–435 (1959)

16.58 I. Lehiste: *Suprasegmentals* (MIT Press, Cambridge 1970)

16.59 O. Jespersen: *Lehrbuch der Phonetik* (Teubner, Leipzig 1926)

16.60 T. Bartholomew: A physical definition of good voice quality in the male voice, J. Acoust. Soc. Am. **6**, 25–33 (1934)

16.61 J. Sundberg: Production and function of the singing formant. In: *Report of the eleventh congress Copenhagen 1972 (Proceedings of the 11th international congress of musicology)*, ed. by H. Glahn, S. Sörensen, P. Ryom (Wilhelm Hansen, Copenhagen 1972) pp. 679–686

16.62 J. Sundberg: Articulatory interpretation of the 'singing formant', J. Acoust. Soc. Am. **55**, 838–844 (1974)

16.63 J. Sundberg: Level and center frequency of the singer´s formant, J. Voice. **15**, 176–186 (2001)

16.64 G. Berndtsson, J. Sundberg: Perceptual significance of the center frequency of the singers formant, Scand. J. Logopedics Phoniatrics **20**, 35–41 (1995)

16.65 L. Dmitriev, A. Kiselev: Relationship between the formant structure of different types of singing voices and the dimension of supraglottal cavities, Fol. Phoniat. **31**, 238–41 (1979)

16.66 P. Ladefoged: *Three areas of experimental phonetics* (Oxford Univ. Press, London 1967)

16.67 J. Barnes, P. Davis, J. Oates, J. Chapman: The relationship between professional operatic soprano voice and high range spectral energy, J. Acoust. Soc. Am. **116**, 530–538 (2004)

16.68 M. Nordenberg, J. Sundberg: Effect on LTAS on vocal loudness variation, Logopedics Phoniatrics Vocology **29**, 183–191 (2004)

16.69 R. Weiss, W.S. Brown, J. Morris: Singer's formant in sopranos: Fact or fiction, J. Voice **15**, 457–468 (2001)

16.70 J.M. Heinz, K.N. Stevens: On the relations between lateral cineradiographs, area functions, and acoustics of speech. In: *Proc. Fourth Int. Congress on Acoustics*, Vol. 1a (1965), paper A44

16.71 C. Johansson, J. Sundberg, H. Willbrand: X-ray study of articulation and formant frequencies in two female singers. In: *Proc. of Stockholm Music Acoustics Conference 1983 (SMAC 83)*, Vol. 46(1), ed. by A. Askenfelt, S. Felicetti, E. Jansson, J Sundberg (Kgl. Musikaliska Akad., Stockholm 1985) pp. 203–218

16.72 T. Baer, J.C. Gore, L.C. Gracco, P. Nye: Analysis of vocal tract shape and dimensions using magnetic

Amazon EX1014 Page 716

resonance imaging: Vowels, J. Acoust. Soc. Am. **90**(2), 799–828 (1991)
16.73 D. Demolin, M. George, V. Lecuit, T. Metens, A. Soquet: Détermination par IRM de l'ouverture du velum des voyelles nasales du français. In: *Actes des XXièmes Journées d'Études sur la Parole* (1996)
16.74 A. Foldvik, K. Kristiansen, J. Kvaerness, A. Torp, H. Torp: *Three-dimensional ultrasound and magnetic resonance imaging: a new dimension in phonetic research* (Proc. Fut. Congress Phonetic Science Stockholm Univ., Stockholm 1995), Vol. 4, 46–49
16.75 B.H. Story, I.R. Titze, E.A. Hoffman: Vocal tract area functions from magnetic resonance imaging, J. Acoust. Soc. Am. **100**, 537–554 (1996)
16.76 O. Engwall: *Talking tongues, D.Sc. thesis* (Royal Institute of Technology (KTH), Stockholm 2002)
16.77 B. Lindblom, J. Sundberg: Acoustical consequences of lip, tongue, jaw and larynx movement, J. Acoust. Soc. Am. **50**, 1166–1179 (1971), also in Papers in Speech Communication: Speech Production, ed. by R.D. Kent, B.S. Atal, J.L. Miller (Acoust. Soc. Am., New York 1991) pp.329–342
16.78 J. Stark, B. Lindblom, J. Sundberg: APEX – an articulatory synthesis model for experimental and computational studies of speech production. In: *Fonetik 96: Papers presented at the Swedish Phonetics Conference TMH-QPSR 2/1996* (Royal Institute of Technology, Stockholm 1996) pp. 45–48
16.79 J. Stark, C. Ericsdotter, B. Lindblom, J. Sundberg: Using X-ray data to calibrate the APEX the synthesis. In: *Fonetik 98: Papers presented at the Swedish Phonetics Conference* (Stockholm Univ., Stockholm 1998)
16.80 J. Stark, C. Ericsdotter, P. Branderud, J. Sundberg, H.-J. Lundberg, J. Lander: The APEX model as a tool in the specification of speaker-specific articulatory behavior. In: *Proc. 14th Int. Congress of the Phonetic Sciences*, ed. by J.J. Ohala (1999)
16.81 C. Ericsdotter: Articulatory copy synthesis: Acoustic performane of an MRI and X-ray-based framework. In: *Proc. 15th Int. Congress of the Phonetic Sciences*, ed. by D. Recasens, M. Josep Solé, J. Romero (Universitat Autònoma de Barcelona, Barcelona 2003), CD-ROM
16.82 C. Ericsdotter: *Articulatory-acoustic relationships in Swedish vowel sounds, PhD thesis* (Stockholm University, Stockholm 2005)
16.83 K.N. Stevens, A.S. House: Development of a quantitative description of vowel articulation, J. Acoust. Soc. Am. **27**, 484–493 (1955)
16.84 S. Maeda: Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: *Speech Production and Speech Modeling*, ed. by W.J. Hardcastle, A. Marchal (Dordrecht, Kluwer 1990) pp. 131–150
16.85 P. Branderud, H. Lundberg, J. Lander, H. Djamshidpey, I. Wäneland, D. Krull, B. Lindblom: X-ray analyses of speech: methodological aspects. In: *Proc. XIIIth Swedish Phonetics Conf. (FONETIK 1998)* (KTH, Stockholm 1998)
16.86 C.Y. Espy-Wilson: Articulatory strategies, speech acoustics and variability. In: *From sound to Sense: 50+ Years of Discoveries in Speech Communication*, ed. by J. Slifka, S. Manuel, M. Mathies (MIT, Cambridge 2004)
16.87 J. Sundberg: Formant technique in a professional female singer, Acustica **32**, 89–96 (1975)
16.88 J. Sundberg, J. Skoog: Dependence of jaw opening on pitch and vowel in singers, J. Voice **11**, 301–306 (1997)
16.89 G. Fant: Glottal flow, models and interaction, J. Phon. **14**, 393–399 (1986)
16.90 E. Joliveau, J. Smith, J. Wolfe: Vocal tract resonances in singing: The soprano voice, J. Acoust. Soc. Am. **116**, 2434–2439 (2004)
16.91 R.K. Potter, A.G. Kopp, H.C. Green: *Visible Speech* (Van Norstrand, New York 1947)
16.92 M. Joosg: Acoustic phonetics, Language **24**, 447–460 (2003), supplement 2
16.93 C.F. Hockett: *A Manual of Phonology* (Indiana Univ. Publ., Bloomington 1955)
16.94 F.H. Guenther: Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production, Psychol. Rev. **102**, 594–621 (1995)
16.95 R.D. Kent, B.S. Atal, J.L. Miller: *Papers in Speech Communication: Speech Perception* (Acoust. Soc. Am., New York 1991)
16.96 S.D. Goldinger, D.B. Pisoni, P. Luce: Speech perception and spoken word recognition. In: *Principles of experimental phonetics*, ed. by N.J. Lass (Mosby, St Louis 1996) pp. 277–327
16.97 H.M. Sussman, D. Fruchter, J. Hilbert, J. Sirosh: Linear correlates in the speech signal: The orderly output constraint, Behav. Brain Sci. **21**, 241–299 (1998)
16.98 B. Lindblom: Economy of speech gestures. In: *The Production of Speech*, ed. by P.F. MacNeilage (Springer, New York 1983) pp. 217–245
16.99 P.A. Keating, B. Lindblom, J. Lubker, J. Kreiman: Variability in jaw height for segments in English and Swedish VCVs, J. Phonetics **22**, 407–422 (1994)
16.100 K. Rapp: A study of syllable timing. In: *STL/Quart. Prog. Status Rep. 1* (Royal Inst. of Technology, Stockholm(1971) pp. 14–19
16.101 F. Koopmans-van Beinum, J. Van der Stelt (Eds.): *Early stages in the development of speech movements* (Stockton, New York 1986)
16.102 K. Oller: Metaphonology and infant vocalizations. In: *Precursors of early speech*, ed. by B. Lindblom, R. Zetterström (Stockton, New York 1986) pp. 21–36

Amazon EX1014 Page 717

16.103 L. Roug, L. Landberg, L. Lundberg: Phonetic development in early infancy, J. Child Language **16**, 19–40 (1989)

16.104 R. Stark: Stages of speech development in the first year of life. In: *Child Phonology: Volume 1: Production*, ed. by G. Yeni-Komshian, J. Kavanagh, C. Ferguson (Academic, New York 1980) pp. 73–90

16.105 R. Stark: Prespeech segmental feature development. In: *Language Acquisition*, ed. by P. Fletcher, M. Garman (Cambridge UP, New York 1986) pp. 149–173

16.106 D.K. Oller, R.E. Eilers: The role of audition in infant babbling, Child Devel. **59**(2), 441–449 (1988)

16.107 C. Stoel-Gammon, D. Otomo: Babbling development of hearing-impaired and normally hearing subjects, J. Speech Hear. Dis. **51**, 33–41 (1986)

16.108 R.E. Eilers, D.K. Oller: Infant vocalizations and the early diagnosis of severe hearing impairment, J. Pediatr. **124**(2), 99–203 (1994)

16.109 D. Ertmer, J. Mellon: Beginning to talk at 20 months: Early vocal development in a young cochlear implant recipient, J. Speech Lang. Hear. Res. **44**, 192–206 (2001)

16.110 R.D. Kent, M.J. Osberger, R. Netsell, C.G. Hustedde: Phonetic development in identical twins who differ in auditory function, J. Speech Hear. Dis. **52**, 64–75 (1991)

16.111 M. Lynch, D. Oller, M. Steffens: Development of speech-like vocalizations in a child with congenital absence of cochleas: The case of total deafness, Appl. Psychol. **10**, 315–333 (1989)

16.112 C. Stoel-Gammon: Prelinguistic vocalizations of hearing-impaired and normally hearing subjects: A comparison of consonantal inventories, J. Speech Hear. Dis. **53**, 302–315 (1988)

16.113 P.F. MacNeilage, B.L. Davis: Acquisition of speech production: The achievement of segmental independence. In: *Speech production and speech modeling*, ed. by W.J. Hardcastle, A. Marchal (Dordrecht, Kluwer 1990) pp. 55–68

16.114 T. Houtgast, H.J.M. Steeneken: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria, J. Acoust. Soc. Am. **77**, 1069–1077 (1985)

16.115 T. Houtgast, H.J.M. Steeneken: *Past, Present and Future of the Speech Transmission Index*, ed. by S.J. van Wijngaarden (NTO Human Factors, Soesterberg 2002)

16.116 R. Drullman, J.M. Festen, R. Plomp: Effect of temporal envelope smearing on speech reception, J. Acoust. Soc. Am. **95**, 1053–1064 (1994)

16.117 R. Drullman, J.M. Festen, R. Plomp: Effect of reducing slow temporal modulations on speech reception, J. Acoust. Soc. Am. **95**, 2670–2680 (1994)

16.118 J. Morton, S. Marcus, C. Frankish: Perceptual centers (P-centers), Psych. Rev. **83**, 405–408 (1976)

16.119 S. Marcus: Acoustic determinants of perceptual center (P-center) location, Perception & Psychophysics **30**, 247–256 (1981)

16.120 G. Allen: The place of rhythm in a theory of language, UCLA Working Papers **10**, 60–84 (1968)

16.121 G. Allen: The location of rhythmic stress beats in English: An experimental study, UCLA Working Papers **14**, 80–132 (1970)

16.122 J. Eggermont: Location of the syllable beat in routine scansion recitations of a Dutch poem, IPO Annu. Prog. Rep. **4**, 60–69 (1969)

16.123 V.A. Kozhevnikov, L.A. Chistovich: Speech Articulation and Perception, JPRS **30**, 543 (1965)

16.124 C.E. Hoequist: The perceptual center and rhythm categories, Lang. Speech **26**, 367–376 (1983)

16.125 K.J. deJong: The correlation of P-center adjustments with articulatory and acoustic events, Perception Psychophys. **56**, 447–460 (1994)

16.126 A.D. Patel, A. Löfqvist, W. Naito: The acoustics and kinematics of regularly timed speech: a database and method for the study of the P-center problem. In: *Proc. 14th Int. Congress of the Phonetic Sciences*, ed. by J.J. Ohala (1999)

16.127 P. Howell: Prediction of P-centre location from the distribution of energy in the amplitude envelope: I & II, Perception Psychophys. **43**, 90–93, 99 (1988)

16.128 B. Pompino-Marschall: On the psychoacoustic nature of the Pcenter phenomenon, J. Phonetics **17**, 175–192 (1989)

16.129 C.A. Harsin: Perceptual-center modeling is affected by including acoustic rate-of-change modulations, Perception Psychophys. **59**, 243–251 (1997)

16.130 C.A. Fowler: Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet, J. Exp. Psychol. Gen. **112**, 386–412 (1983)

16.131 H. Fujisaki: Dynamic characteristics of voice fundamental frequency in speech and singing. In: *The Production of Speech*, ed. by P.F. MacNeilage (Springer, New York 1983) pp. 39–55

16.132 J. Frid: *Lexical and acoustic modelling of Swedish prosody, Dissertation* (Lund University, Lund 2003)

16.133 S. Öhman: Numerical model of coarticulation, J. Acoust. Soc. Am. **41**, 310–320 (1967)

16.134 J. t'Hart: F0 stylization in speech: Straight lines versus parabolas, J. Acoust. Soc. Am. **90**(6), 3368–3370 (1991)

16.135 D. Abercrombie: *Elements of General Phonetics* (Edinburgh Univ. Press, Edinburgh 1967)

16.136 K.L. Pike: *The intonation of America English* (Univ. of Michigan Press, Ann Arbor 1945)

16.137 G. Fant, A. Kruckenberg: Notes on stress and word accent in Swedish. In: *STL/Quart. Prog. Status Rep. 2-3* (Royal Inst. of Technology, Stockholm 1994) pp. 125–144

16.138 R. Dauer: Stress timing and syllable-timing reanalyzed, J. Phonetics **11**, 51–62 (1983)

Amazon EX1014 Page 718

16.139 A. Eriksson: *Aspects of Swedish rhythm, PhD thesis, Gothenburg Monographs in Linguistics* (Gothenburg University, Gothenburg 1991)
16.140 O. Engstrand, D. Krull: Duration of syllable-sized units in casual and elaborated speech: cross-language observations on Swedish and Spanish, TMH-QPSR **44**, 69–72 (2002)
16.141 A.D. Patel, J.R. Daniele: An empirical comparison of rhythm in language and music, Cognition **87**, B35–B45 (2003)
16.142 D. Huron, J. Ollen: Agogic contrast in French and English themes: Further support for Patel and Daniele (2003), Music Perception **21**, 267–271 (2003)
16.143 D.H. Klatt: Synthesis by rule of segmental durations in English sentences. In: *Frontiers of speech communication research*, ed. by B. Lindblom, S. Öhman (Academic, London 1979) pp. 287–299
16.144 B. Lindblom: Final lengthening in speech and music. In: *Nordic Prosody*, ed. by E. Gårding, R. Bannert (Department of Linguistics Lund University, Lund 1978) pp. 85–102
16.145 A. Friberg, U Battel: Structural communication. In: *The Science and Psychology of Music Performance*, ed. by R. Parncutt, GE McPherson (Oxford Univ., Oxford 2001) pp. 199–218
16.146 J. Sundberg: Emotive transforms, Phonetica **57**, 95–112 (2000)
16.147 Brownlee: *The role of sentence stress in vowel reduction and formant undershoot: A study of lab speech and informal spontaneous speech, PhD thesis* (University of Texas, Austin 1996)
16.148 S.-J. Moon: *An acoustic and perceptual study of undershoot in clear and citation- form speech, PhD dissertation* (Univ. of Texas, Austin 1991)
16.149 K.N. Stevens, A.S. House: Perturbation of vowel articulations by consonantal context. An acoustical study, JSHR **6**, 111–128 (1963)
16.150 B. Lindblom: Spectrographic study of vowel reduction, J. Acoust. Soc. Am. **35**, 1773–1781 (1963)
16.151 P. Delattre: An acoustic and articulatory study of vowel reduction in four languages, IRAL-Int. Ref. Appl. VII/ **4**, 295–325 (1969)
16.152 D.P. Kuehn, K.L. Moll: A cineradiographic study of VC and CV articulatory velocities, J. Phonetics **4**, 303–320 (1976)
16.153 J.E. Flege: Effects of speaking rate on tongue position and velocity of movement in vowel production, J. Acoust. Soc. Am. **84**(3), 901–916 (1988)
16.154 R.J.J.H. van Son, L.C.W. Pols: "Formant movements of Dutch vowels in a text, read at normal and fast rate, J. Acoust. Soc. Am. **92**(1), 121–127 (1992)
16.155 D. van Bergem: *Acoustic and Lexical Vowel Reduction, Doctoral Dissertation* (University of Amsterdam, Amsterdam 1995)
16.156 W.L. Nelson, J.S. Perkell, J.R. Westbury: Mandible movements during increasingly rapid articulations of single syllables: Preliminary observations, J. Acoust. Soc. Am. **75**(3), 945–951 (1984)
16.157 S.-J. Moon, B. Lindblom: Interaction between duration, context and speaking style in English stressed vowels, J. Acoust. Soc. Am. **96**(1), 40–55 (1994)
16.158 C.S. Sherrington: *Man on his nature* (MacMillan, London 1986)
16.159 R. Granit: *The Purposive Brain* (MIT, Cambridge 1979)
16.160 N. Bernstein: *The coordination and regulation of movements* (Pergamon, Oxford 1967)
16.161 P.F. MacNeilage: Motor control of serial ordering of speech, Psychol. Rev. **77**, 182–196 (1970)
16.162 A. Löfqvist: Theories and Models of Speech Production. In: *The Handbook of Phonetic Sciences*, ed. by W.J. Hardcastle, J. Laver (Blackwell, Oxford 1997) pp. 405–426
16.163 J.S. Perkell: Articulatory processes. In: *The Handbook of Phonetic Sciences. 5*, ed. by W.J. Hardcastle, J. Laver. (Blackwell, Oxford 1997) pp. 333–370
16.164 J. Sundberg, R. Leandersson, C. von Euler, E. Knutsson: Influence of body posture and lung volume on subglottal pressure control during singing, J. Voice **5**, 283–291 (1991)
16.165 T. Sears, J. Newsom Davis: The control of respiratory muscles during voluntary breathing. In: *Sound production in man*, ed. by A. Bouhuys et al. (Annals of the New York Academy of Science, New York 1968) pp. 183–190
16.166 B. Lindblom, J. Lubker, T. Gay: Formant frequencies of some fixed-mandible vowels and a model of motor programming by predictive simulation, J. Phonetics **7**, 147–161 (1979)
16.167 T. Gay, B. Lindblom, J. Lubker: Production of bite-block vowels: Acoustic equivalence by selective compensation, J. Acoust. Soc. Am. **69**(3), 802–810 (1981)
16.168 W.J. Hardcastle, J. Laver (Eds.): *The Handbook of Phonetic Sciences* (Blackwell, Oxford 1997)
16.169 J. S. Perkell, D. H. Klatt: *Invariance and variability in speech processes* (LEA, Hillsdale 1986)
16.170 A. Liberman, I. Mattingly: The motor theory of speech perception revised, Cognition **21**, 1–36 (1985)
16.171 C.A. Fowler: An event approach to the study of speech perception from a direct- realist perspective, J. Phon. **14**(1), 3–28 (1986)
16.172 E.L. Saltzman, K.G. Munhall: A dynamical approach to gestural patterning in speech production, Ecol. Psychol. **1**, 91–163 (1989)
16.173 M. Studdert-Kennedy: How did language go discrete?. In: *Evolutionary Prerequisites of Language*, ed. by M. Tallerman (Oxford Univ., Oxford 2005) pp. 47–68
16.174 R. Jakobson, G. Fant, M. Halle: *Preliminaries to Speech Analysis, Acoustics Laboratory, MIT Tech. Rep. No. 13* (MIT, Cambridge 1952)
16.175 B. Lindblom: Explaining phonetic variation: A sketch of the H&H theory. In: *Speech Produc-*

Amazon EX1014 Page 719

tion and Speech Modeling, ed. by W.J. Hardcastle, A. Marchal (Dordrecht, Kluwer 1990) pp. 403–439

16.176 B. Lindblom: Role of articulation in speech perception: Clues from production, J. Acoust. Soc. Am. **99**(3), 1683–1692 (1996)

16.177 E. Rapoport: Emotional expression code in opera and lied singing, J. New Music Res. **25**, 109–149 (1996)

16.178 J. Sundberg, E. Prame, J. Iwarsson: Replicability and accuracy of pitch patterns in professional singers. In: *Vocal Fold Physiology, Controlling Complexity and Chaos*, ed. by P. Davis, N. Fletcher (Singular, San Diego 1996) pp. 291–306, Chap. 20

16.179 J.J. Ohala: An ethological perspective on common cross-language utilization of F0 of voice, Phonetica **41**, 1–16 (1984)

16.180 I. Fónagy: Hörbare Mimik, Phonetica **1**, 25–35 (1967)

16.181 K. Scherer: Expression of emotion in voice and music, J. Voice **9**, 235–248 (1995)

16.182 P. Juslin, P. Laukka: Communication of emotions in vocal expression and music performance: Different channels, same code?, Psychol. Rev. **129**, 770–814 (2003)

16.183 J. Sundberg, J. Iwarsson, H. Hagegård: A singers expression of emotions in sung performance,. In: *Vocal Fold Physiology: Voice Quality Control*, ed. by O. Fujimura, M. Hirano (Singular, San Diego 1995) pp. 217–229

Amazon EX1014 Page 720

17. Computer Music

This chapter covers algorithms, technologies, computer languages, and systems for computer music. Computer music involves the application of computers and other digital/electronic technologies to music composition, performance, theory, history, and perception. The field combines digital signal processing, computational algorithms, computer languages, hardware and software systems, acoustics, psychoacoustics (low-level perception of sounds from the raw acoustic signal), and music cognition (higher-level perception of musical style, form, emotion, etc.). Although most people would think that analog synthesizers and electronic music substantially predate the use of computers in music, many experiments and complete computer music systems were being constructed and used as early as the 1950s.

Because of this rich legacy, and the large number of researchers working on digital audio (primarily in speech research laboratories), there are a large number of algorithms for synthesizing sound using computers. Thus, a significant emphasis in this chapter will be placed on digital sound synthesis and processing, first providing an overview of the representation of audio in digital systems, then covering most of the popular algorithms for digital analysis and synthesis of sound.

17.1 Computer Audio Basics 714
17.2 Pulse Code Modulation Synthesis 717
17.3 Additive (Fourier, Sinusoidal) Synthesis . 719
17.4 Modal (Damped Sinusoidal) Synthesis 722
17.5 Subtractive (Source-Filter) Synthesis..... 724
17.6 Frequency Modulation (FM) Synthesis 727
17.7 FOFs, Wavelets, and Grains..... 728
17.8 Physical Modeling (The Wave Equation) . 730
17.9 Music Description and Control..... 735
17.10 Composition..... 737
17.11 Controllers and Performance Systems 737
17.12 Music Understanding and Modeling by Computer 738
17.13 Conclusions, and the Future 740
References 740

Pulse code modulation (PCM) is the means for sampling and retrieval of audio in computers, and PCM synthesis uses combinations of prerecorded waveforms to reconstruct speech, sound effects, and music instrument sounds, based largely on the work of Joseph Fourier, who gave us a technique for formulating any waveform from sine waves. Additive synthesis uses fundamental waveforms (often sine waves), to construct more complicated waves. PCM and sinusoidal (Fourier) additive synthesis are often called *nonparametric* techniques. The word parametric in this context means that an algorithm with a few (but hopefully expressive) parameters can be used to generate a large variety of waveforms and spectral properties by varying the parameters.

Since PCM involves recording and playing back waveforms, no actual expressive parameters are available in PCM systems. Similarly, Fourier's theorem says that we can represent any waveform with a sum of sine waves, but in fact it might take a huge number of sine waves to represent a particular waveform accurately. So, without further work to *parameterize* PCM samples or large groups of additive sine waves, these techniques are nonparametric.

Modal synthesis recognizes the fact that many vibrating systems exhibit a relatively few strong sinusoidal modes (natural frequencies) of vibration. Rigid bars and plates, plucked strings, and other structures are good candidates for modal synthesis, where after being excited (struck, plucked, etc.) the vibration/sound generation is restricted to a limited set of exponentially decaying sine waves. Modal synthesis is considered by some to be parametric, since the number of sine waves needed is often quite limited. Others claim that modal

Amazon EX1014 Page 721

synthesis is nonparametric, since there is no small set of meaningful and expressive parameters that control variation in the sound.

Subtractive synthesis begins with spectrally complex waveforms (such as pulses or noise), and uses filters to shape the spectrum and waveform to the desired result. Subtractive synthesis works particularly well for certain classes of sounds, such as human speech, where a spectrally rich source (pulses of the vocal folds) is filtered by a time-varying filter (the acoustic tube of the vocal tract). Subtractive synthesis is parametric, because only a few descriptive numbers control such a model. For example, a subtractive speech model might be controlled by numbers expressing how pitched versus noisy the voice is, the voice pitch, the loudness, and 8–10 locations of the resonances of the vocal tract. Using only a dozen or so variable numbers (parameters), a rich variety of vowel and consonant sounds can be synthesized.

Frequency modulation (FM) synthesis exploits properties of nonlinearity to generate complex spectra and waveforms from much simpler ones. Modulating the frequency or phase of a sine wave (the carrier) by another sine wave (the modulator) causes multiple *sideband* sinusoids to appear at frequencies related to the ratio of the carrier and modulator frequencies. Thus an important parameter in FM is the carrier:modulator (C:M) ratio, and another is the amount (index) of modulation, which when increased causes the number of sidebands to increase.

Wavelets are waveforms that are usually compact in time and frequency. They are viewed by some as related to the fundamental unit of sound, by others as convenient building blocks for constructing sonic textures, and by others as related to important small events in nature and acoustics. *Formes d'onde formantiques* (FOFs, French for formant wave functions) are special wavelets that model the individual resonances of the vocal tract, and thus have been successfully used for modeling speech and singing.

Physical models solve the wave equation in space and time to generate waveforms in the same ways that physical acoustics does. Efficient techniques combining discrete simulation of physics with modern digital filtering techniques allow models of plucked strings, wind instruments, percussion, the vocal tract, and many other musical systems to be solved efficiently by computers.

After the presentation and examples of the main synthesis algorithms, some industry standards for representing and manipulating musical data in computerized systems will be discussed. The musical instrument digital interface (MIDI), downloadable sounds (DLS), open sound control (OSC), structured audio orchestra language (SAOL), and other score and control standards will be briefly covered.

Some languages and systems for computer music composition are discussed, including the historical MusicX languages, and many modern languages such as Cmusic and MAX/MSP, followed by a few real-time performance and interaction systems. Finally, some aspects of computer modeling of human listening and systems for machine-assisted and automatic composition are covered.

17.1 Computer Audio Basics

Typically, digital audio signals are formed by *sampling* analog (continuous in time and amplitude) signals at regular intervals in time, and then *quantizing* the amplitudes to discrete values. The process of sampling a waveform, holding the value, and quantizing the value to the nearest number that can be digitally represented (as a specific integer on a finite range of integers) is called analog-to-digital (A to D, or A/D) conversion [17.1]. A device that does A/D conversion is called an analog-to-digital converter (ADC). Coding and representing waveforms in sampled digital form is called pulse code modulation (PCM), and digital audio signals are often called PCM audio. The process of converting a sampled signal back into an analog signal is called digital-to-analog conversion (D to A, or D/A), and the device is called a digital-to-analog converter (DAC). Low-pass filtering (smoothing the samples to remove unwanted high frequencies) is necessary to reconstruct the sampled signal back into a smooth continuous-time analog signal. This filtering is usually contained in the DAC hardware.

The time between successive samples is usually denoted by T. Sampling an analog signal first requires filtering it to remove unwanted high frequencies (more on this shortly), holding the value steady for a period while a stable measurement can be made, then associating the analog value with a digital number (coding). Analog signals can have any of the infinity of real-numbered amplitude values. Since computers work with fixed word sizes (8 bit bytes, 16 bit words, etc.), digital signals can only have a finite number of amplitude val-

Amazon EX1014 Page 722

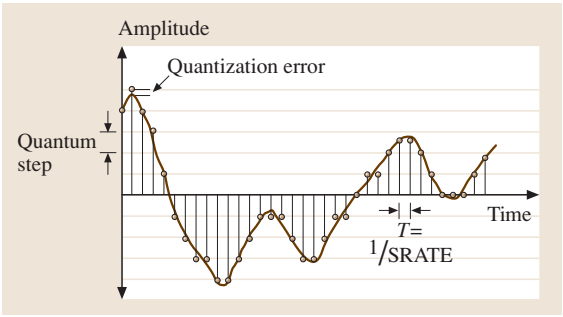


Fig. 17.1 Linear sampling and quantization

ues. In converting from analog to digital, rounding takes place and a given analog value must be quantized to the nearest digital value. The difference between quantization steps is called the quantum (not as in quantum physics or leaps, but the Latin word for a fixed-sized jump in value or magnitude). Sampling and quantization is shown in Fig. 17.1; note the errors introduced in some sample values due to the quantization process.

A fundamental law of digital signal processing states that, if an analog signal is bandlimited with bandwidth B Hz, (Hz = Hertz = samples per second), the signal can be periodically sampled at a rate of $2B$ Hz or greater, and exactly reconstructed from the samples. Bandlimited with bandwidth B means that no frequencies above B exist in the signal. The rate $2B$ is called the sampling rate, and B is called the Nyquist frequency. Intuitively, a sine wave at the highest frequency B present in a bandlimited signal can be represented using two samples per period (one sample at each of the positive and negative peaks), corresponding to a sampling frequency of $2B$. All signal components at lower frequencies can be uniquely represented and distinguished from each other using this same sampling frequency of $2B$. If there are components present in a signal at frequencies greater than $1/2$ the sampling rate, these components will not be represented properly, and will *alias* (i. e.,, show up as frequencies different from their original values).

To avoid aliasing, ADC hardware often includes filters that limit the bandwidth of the incoming signal before sampling takes place, automatically changing as a function of the selected sampling rate. Figure 17.2 shows aliasing in complex and simple (sinusoidal) waveforms. Note the loss of detail in the complex waveform. Also note that samples at less than two times the frequency of a sine wave could also have arisen from a sine wave of much lower frequency. This is the fundamental nature of aliasing, because frequencies higher than $1/2$ sample rate *alias* as lower frequencies.

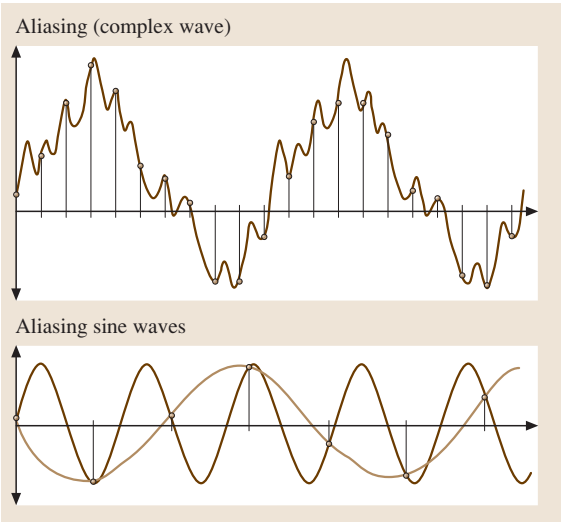


Fig. 17.2 Because of inadequate sampling rate, aliasing causes important features to be lost

Humans can perceive frequencies from roughly 20 Hz to 20 kHz, thus requiring a minimum sampling rate of at least 40 kHz. Speech signals are often sampled at 8 kHz (*telephone quality*) or 11.025 kHz, while music is usually sampled at 22.05 kHz, 44.1 kHz (the sampling rate used on audio compact discs), or 48 kHz. Some new formats allow for sampling rates of 96 kHz, and even 192 kHz. This is because some engineers believe that we can actually hear things, or the effects of things, higher than 20 kHz. Dolphins, dogs, and some other animals can actually hear higher than us, so these new formats could be considered as catering to those potential markets.

In a digital system, a fixed number of binary digits (bits) are used to sample the analog waveform, by quantizing it to the closest number that can be represented. This quantization is accomplished either by rounding to the quantum value nearest the actual analog value, or by truncation to the nearest quantum value less than or equal to the actual analog value. With uniform sampling in time, a properly bandlimited signal can be exactly recovered provided that the sampling rate is twice the bandwidth or greater, but only if there is no quantization. When the signal values are rounded or truncated, the amplitude difference between the original signal and the quantized signal is lost forever. This can be viewed as an additive noise component upon reconstruction. Using the additive-noise assumption gives an approximate best-case signal-to-quantization-noise ratio (SNR) of approximately $6N$ dB, where N is the number of

Amazon EX1014 Page 723

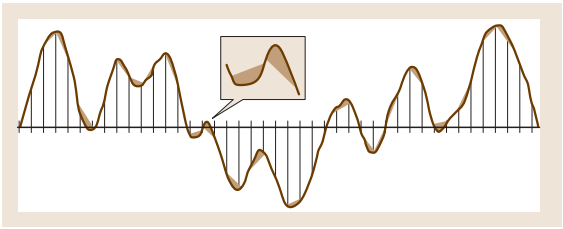


Fig. 17.3 Resampling by linear interpolation. Grey regions show the errors due to linear interpolation

bits. Using this approximation implies that a 16 bit linear quantization system will exhibit an SNR of approximately 96 dB. 8 bit quantization exhibits a signal-to-quantization-noise ratio of approximately 48 dB. Each extra bit improves the signal-to-noise ratio by about 6 dB.

Most computer audio systems use two or three types of audio data words. 16 bit (per channel) data is quite common, and this is the data format used in compact disc systems. Many recent (high-definition) formats allow for 24 bit samples. 8 bit data is common for speech data in personal computer (PC) and telephone systems, usually using methods of quantization that are nonlinear. In such systems the quantum is smaller for small amplitudes, and larger for large amplitudes.

Changing the playback sample rate on sampled sound results in a pitch shift. Many systems for recording, playback, processing, and synthesis of music, speech, or other sounds allow or require flexible control of pitch (sample rate). The most accurate pitch control is necessary for music synthesis. In sampling synthesis, this is accomplished by dynamic sample-rate conversion (interpolation). In order to convert a sampled data signal from one sampling rate to another, three steps are required: bandlimiting, interpolation, and resampling.

Bandlimiting: First, it must be assured that the original signal was properly bandlimited to less than half the new sampling rate. This is always true when *upsampling* (i. e., converting from a lower sampling rate to a higher one). When *downsampling*, however, the signal must be first low-pass filtered to exclude frequencies greater than half the new target rate.

Interpolation: When resampling by any but simple integer factors such as 2, 4, 1/2, etc., the task requires the computation of the values of the original bandlimited analog signal at the correct points between the existing sampled points. This is called interpolation. The most common form of interpolation used is linear interpolation, where the fractional time samples needed are computed as a linear combination of the two surrounding samples. Many assume that linear interpolation is the correct means, or at least an adequate means for accomplishing audio sample rate conversion. Figure 17.3 shows resampling by linear interpolation; note the errors shown by the shaded regions.

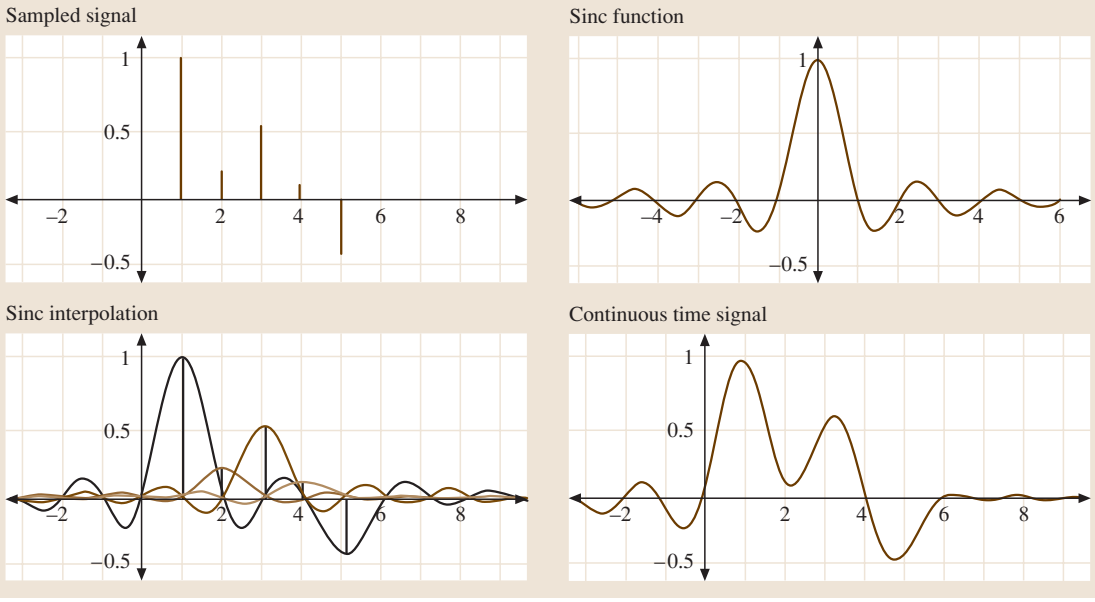


Fig. 17.4 Sinc interpolation for reconstruction and resampling

Some might notice that linear interpolation is not quite adequate, so they might assume that the correct solution must lie in more elaborate curve-fitting techniques using quadratics, cubics, or higher-order splines, and indeed these types of interpolation can be adequate for some applications. To arrive at the correct answer (provably correct from theory) the interpolation task should be viewed and accomplished as a filtering problem, with the filter designed to meet some appropriate error criterion. Linear time-invariant filtering is accomplished by convolution with a filter function. If the resampling filter is defined appropriately, we can exactly reconstruct the original analog waveform from the samples.

The correct (ideal in a provable digital signal processing sense) way to perform interpolation is convolution with the sinc function, defined as:

$$\mathrm{sinc}(t/T) = \sin(\pi t/T)/(\pi t/T)\,,$$

where

$$T = 1/\mathrm{SRATE}\,.$$

The sinc function is the ideal low-pass filter with a cutoff of SRATE/2, where SRATE is the sampling rate. Figure 17.4 shows reconstruction of a continuous waveform by convolution of a digital signal with the sinc function. Each sample is multiplied by a corresponding continuous sinc function, and those are added up to arrive at the continuous reconstructed signal [17.2].

Resampling: This is usually accomplished at the same time as interpolation, because it is not necessary to reconstruct the entire continuous waveform in order to acquire new discrete samples. The resampling ratio can be time varying, making the problem a little more difficult. However, viewing the problem as a filter-design and implementation issue allows for guaranteed tradeoffs of quality and computational complexity.

17.2 Pulse Code Modulation Synthesis

The majority of digital sound and music synthesis today is accomplished via the playback of stored pulse code modulation (PCM) waveforms. Single-shot playback of entire segments of stored sounds is common for sound effects, narrations, prompts, segments of music, etc. Most high-quality modern electronic music synthesizers, speech synthesis systems, and PC software systems for sound synthesis use pre-stored PCM as the basic data. This data is sometimes manipulated to yield the final output sound(s).

There are a number of different ways to look at sound for computer music, with PCM being only one. We can look at the physics that produce the sound and try to model those. We could also look at the spectrum of the sound and other characteristics having to do with the perception of those sounds. Indeed, much of the legacy of computer music has revolved around parametric (using mathematical algorithms, controlled by a few well-chosen/-designed control parameters) analysis and synthesis algorithms. We will discuss most of the commonly used algorithms later, but first we should look at PCM in more depth.

For speech, the most common synthesis technique is *concatenative* synthesis [17.3]. Concatenative phoneme synthesis relies on the concatenation of roughly 40 pre-stored phonemes (for English). Examples of vowel phonemes are /i/ as in beet, /I/ as in bit, /a/ as in father, etc. Examples of nasals are /m/ as in mom, /n/ as in none, /ng/ as in sing, etc. Examples of fricative consonant phonemes are /s/ as in sit, /sh/ as in ship, /f/ as in fifty, etc. Examples of voiced fricative consonants are /z/, /v/ (visualize), etc. Examples of plosive consonants are /t/ as in tat, /p/ as in pop, /k/ as in kick, etc. Examples of voiced plosives include /d/, /b/, /g/ (dude, bob, gag) etc. Vowels and nasals are essentially periodic pitched sounds, so the minimal required stored waveform is only one single period of each. Consonants require more storage because of their noisy (non-pitched, aperiodic) nature.

The quality of concatenative phoneme synthesis is generally considered quite low, due to the simplistic assumption that all of the pitched sounds (vowels, etc.) are purely periodic. Also, simply *gluing* /s/ /I/ and /ng/ together does not make for a high-quality realistic synthesis of the word "sing". In actual speech, phonemes gradually blend into each other as the jaw, tongue, and other articulators move with time.

Accurately capturing the transitions between phonemes with PCM requires recording transitions from phoneme to phoneme, called *diphones*. A concatenative diphone synthesizer blends together stored diphones. Examples of diphones include see, she, thee, and most of the roughly 40x40 possible combinations of phonemes. Much more storage is necessary for a diphone synthesizer, but the resulting increase in quality is significant. PCM speech synthesis can be improved further by stor-

Amazon EX1014 Page 725

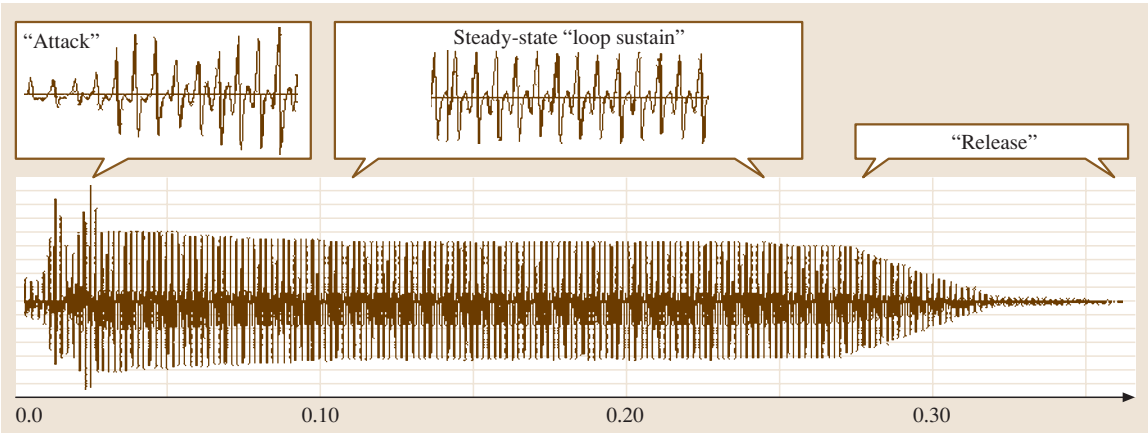


Fig. 17.5 Wave-table synthesis of a trumpet tone

ing multiple samples, for different pitch ranges, genders, voice qualities, etc.

For musical sounds, it is common to store only a loop, or table, of the periodic component of a recorded sound waveform and play that loop back repeatedly. This is called "wave-table synthesis" [17.4]. For more realism, the attack or beginning portion of the recorded sound can be stored in addition to the periodic steady-state part. Figure 17.5 shows the synthesis of a trumpet tone starting with an attack segment, followed by repetition of a periodic loop, ending with an enveloped decay (or release). The *envelope* is a synthesizer/computer music term for a time-varying change applied to a waveform amplitude, or other parameter. Envelopes are often described by four components: the *attack time*, the *decay time* (decay here means the initial decay down to the steady state segment), the *sustain level*, and the *release time* (final decay). Hence, envelopes are sometimes called ADSRs.

Originally called *sampling synthesis* in the music industry, any synthesis using stored PCM waveforms has now become commonly known as "wave-table synthesis". Filters are usually added to high-quality wave-table synthesis, allowing control of spectral brightness as a function of intensity, and to get more variety of sounds from a given set of samples.

Pitch shifting of a PCM sample or wave table is accomplished via interpolation as discussed in the previous section. A given sample can be pitch-shifted only so far in either direction before it begins to sound unnatural. This can be dealt with by storing multiple recordings of the sound at different pitches, and switching or interpolating between these upon resynthesis. This is called *multi-sampling*. Multi-sampling might also include the storage of separate samples for loud and soft sounds. Linear or other interpolation is used to blend the loudness multi-samples as a function of the desired synthesized volume. This adds realism, because loudness is not simply a matter of amplitude or power, and most sound sources exhibit spectral variations as a function of loudness. For example, there is usually more high-frequency energy (*brightness*) in loud sounds than in soft sounds. Filters can also be used to add spectral variation.

A common tool used to describe the various components and steps of signal processing in performing digital music synthesis is the *synthesizer patch* (historically named from hooking various electrical components together using patch cords). In a patch, a set of fairly commonly agreed building blocks, called *unit generators* (also called modules, plug-ins, operators, op-codes, and other terms) are hooked together in a signal flow diagram. This historical [17.5] graphical method of describing signal generation and processing affords a visual representation that is easily printed in pa-

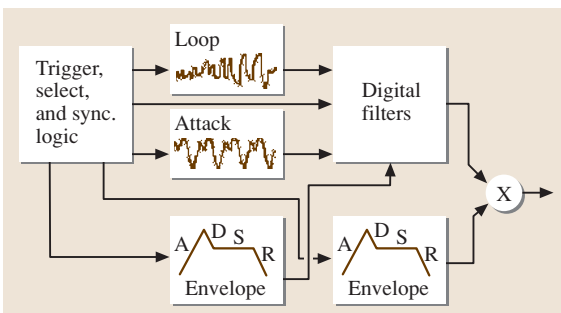


Fig. 17.6 A synthesis *patch* showing interconnection of *unit generators*

pers, textbooks, patents, etc. Further, graphical patching systems and languages have been important to the development and popularization of certain algorithms, and computer music in general. Figure 17.6 shows a PCM synthesizer patch with attack and loop wave tables whose amplitudes are controlled by an envelope generator, and a time-varying filter (also controlled by another envelope generator).

17.3 Additive (Fourier, Sinusoidal) Synthesis

Lots of sound-producing objects and systems exhibit sinusoidal modes. A plucked string might exhibit many modes, with the strength of each mode determined by the boundary conditions of the terminations, and the shape of the excitation pluck. Striking a metal plate with a hammer excites many of the vibrational modes of the plate, determined by the shape of the plate, and by where it is struck. A singing voice, struck drum head, bowed violin string, struck bell, or blown trumpet exhibit oscillations characterized by a sum of sinusoids. The recognition of the fundamental nature of the sinusoid gives rise to a powerful model of sound synthesis based on summing sinusoidal modes.

These modes have a very special relationship in the case of the plucked string, a singing voice, and some other limited systems, in that their frequencies are all integer multiples (at least approximately) of one basic sinusoid, called the *fundamental*. This special series of sinusoids is called a *harmonic series*, and lies at the basis of the Fourier-series representation of shapes, waveforms, oscillations, etc. The Fourier series [17.6] solves many types of problems, including physical problems with boundary constraints, but is also applicable to any shape or function. Any periodic waveform (repeating over and over again) F_{per} can be transformed into a Fourier series, written as:

$$F_{\text{per}}(t) = a_0 + \Sigma_m [b_m \cos(2\pi f_0 m t) + c_m \sin(2\pi f_0 m t)] \,. \tag{17.1}$$

The limits of the summation are technically infinite, but we know that we can cut off our frequencies at the Nyquist frequency for digital signals. The a_0 term is a constant offset, or the average of the waveform. The b_m and c_m coefficients are the weights of the mth harmonic cosine and sine terms. If the function $F_{\text{per}}(t)$ is purely even about $t = 0$ $[(F(-t) = F(t))]$, only cosines are required to represent it, and only the b_m terms would be nonzero. Similarly, if the function $F_{\text{per}}(t)$ is odd $[(F(-t) = -F(t))]$, only the c_m terms would be required. An arbitrary function $F_{\text{per}}(t)$ will require sinusoidal harmonics of arbitrary (but specific) amplitudes and phases. The magnitude and phase of the mth harmonic in the Fourier series can be found by:

$$A_m = \sqrt{b_m^2 + c_m^2} \,, \tag{17.2a}$$

$$\theta_m = \arctan\,(c_m / b_m) \,. \tag{17.2b}$$

Phase is defined relative to the cosine, so if c_m is zero, θ_m is zero. As a brief example, Fig. 17.7 shows the first few sinusoidal harmonics required to build up an approximation of a square wave. Note that due to symmetries (boundary conditions), only odd sine harmonics are required. Using more sines improves the approximation.

The process of solving for the sine and cosine components of a signal or waveform is called Fourier analysis, or the Fourier transform. If the frequency variable is sampled (as is the case in the Fourier series, represented by m), and the time variable t is sampled as well (as it is in PCM waveform data, represented by n), then the Fourier transform is called the *discrete Fourier transform* (DFT). The DFT is given by

$$F(m) = \sum_{n=0}^{N-1} f(n)[\cos(2\pi mn/N) - \mathrm{i}\sin(2\pi mn/N)] \,. \tag{17.3}$$

Where N is the length of the signal being analyzed. The inverse DFT (IDFT) is very similar to the Fourier

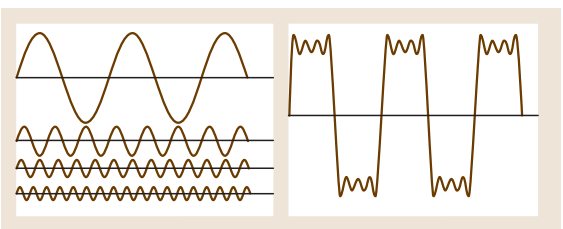


Fig. 17.7 A sum of odd harmonics approximates a square wave

series

$$f(n) = 1/N \sum_{m=0}^{N-1} F(m)[\cos(2\pi mn/N) + \mathrm{i}\sin(2\pi mn/N)] \,. \quad (17.4)$$

The imaginary number $\mathrm{i}(=\sqrt{-1})$ is used to place the cosine and sine components in a unique mathematical arrangement, where odd $[(x(-n) = -x(n))]$ terms of the waveform are represented as imaginary components, and even $[(x(-n) = x(n))]$ terms are represented as real components. This gives us a means to talk about the magnitude and phase in terms of the magnitude and phase of $F(m)$ (a complex number).

There is a near-mystical expression of equality in mathematics known as Euler's identity, which links trigonometry, exponential functions, and complex numbers in a single equation

$$\mathrm{e}^{\mathrm{i}\theta} = \cos(\theta) + \mathrm{i}\sin(\theta) \,.$$

We can use Euler's identity to write the DFT and IDFT in shorthand

$$F(m) = \sum_{n=0}^{N-1} f(n) e^{-\mathrm{i}2\pi mn/N} \,, \quad (17.5)$$

$$f(n) = 1/N \sum_{m=0}^{N-1} F(m)\,\mathrm{e}^{\mathrm{i}2\pi mn/N} \,. \quad (17.6)$$

Converting the cosine/sine form to the complex exponential form allows lots of manipulations that would be difficult otherwise. But we can also write the DFT in real-number terms as a form of the Fourier series

$$f(n) = 1/N \sum_{m=0}^{N-1} F_b(n)\cos(2\pi mn/N) + F_c(n)\sin(2\pi mn/N) \,, \quad (17.7)$$

where

$$F_b(m) = \sum_{n=0}^{N-1} f(n)\cos(2\pi mn/N) \,, \quad (17.8)$$

$$F_c(m) = \sum_{n=0}^{N-1} -f(n)\sin(2\pi mn/N) \,. \quad (17.9)$$

The fast Fourier transform (FFT) is a cheap way of calculating the DFT. There are thousands of references on the FFT [17.6], and scores of implementations of it, so for our purposes we will just say that it is much more efficient than trying to compute the DFT directly from the definition. A well-crafted FFT algorithm for real input data takes on the order of $N\log_2(N)$ multiply–adds to compute. Comparing this to the N^2 multiplies of the DFT, N does not have to be very big before the FFT is a winner. There are some down sides, such as the fact that FFTs can only be computed for signals whose lengths are exactly powers of 2, but the advantages of using it often outweigh the annoying power-of-two problems. Practically speaking, users of FFTs usually carve up signals into small chunks (powers of two), or *zero pad* a signal out to the next biggest power of two.

The short-time Fourier transform (STFT) breaks up the signal and applies the Fourier transform to each segment individually [17.7]. By selecting the window size (length of the segments), and hop size (how far the window is advanced along the signal) to be perceptually relevant, the STFT can be used as an approximation of human audio perception. Figure 17.8 shows the waveform of the utterance of the word "synthesize", and some STFT spectra corresponding to windows at particular points in time.

If we inspect the various spectra in Fig. 17.8, we can note that the vowels exhibit harmonic spectra (clear, evenly spaced peaks corresponding to the harmonics of the pitched voice), while the consonants exhibit noisy spectra (no clear sinusoidal peaks). Recognizing that some sounds are well approximated/modeled by additive sine waves [17.8], while other sounds are essentially noisy, spectral modeling [17.7] breaks the sound into deterministic (sines) and stochastic (noise) components. Figure 17.9 shows a general sines+noise additive synthesis model, allowing us to control the amplitudes and frequencies of a number of sinusoidal oscillators, and model the noisy component with a noise source and a spectral shaping filter.

The beauty of this type of model is that it recognizes the dominant sinusoidal nature of many sounds, while still recognizing the noisy components that might be also present. More efficient and parametric representations, and many interesting modifications can be made to the signal on resynthesis. For example, removing the harmonics from voiced speech, followed by resynthesizing with a scaled version of the noise residual, can result in the synthesis of whispered speech.

One further improvement to spectral modeling is the recognition [17.9] that there are often brief (impulsive) moments in sounds that are really too short in time to be adequately analyzed by spectrum analysis. Further, such moments in the signal usually corrupt the sinusoidal/noise analysis process. Such

Amazon EX1014 Page 728

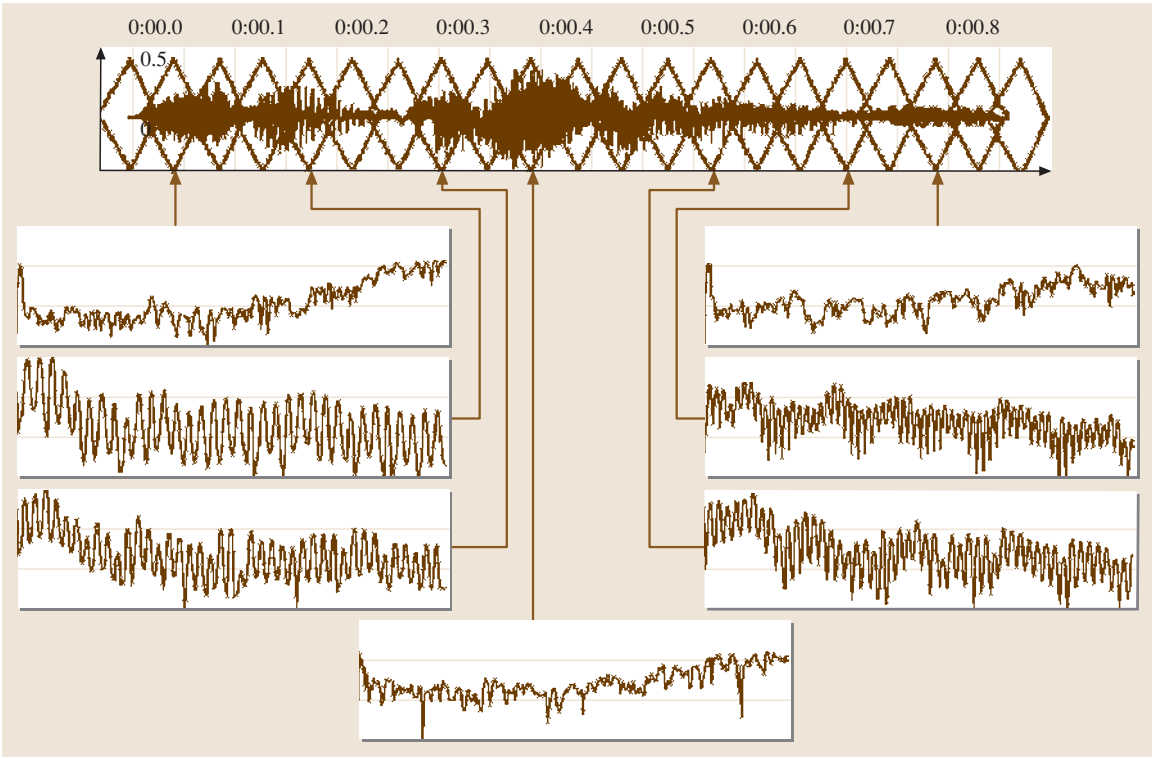


Fig. 17.8 Some STFT frames of the word 'synthesize'

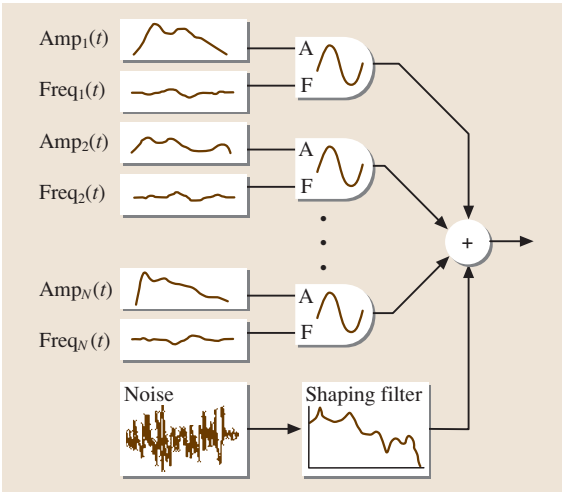


Fig. 17.9 Sinusoidal additive model with filtered noise added for spectral modeling synthesis

events, called transients, can be modeled in other ways (often by simply keeping the stored PCM for that segment).

Using the short-time Fourier transform, the *phase vocoder* (VoiceCoder) [17.10, 11] processes sound by calculating and maintaining both magnitude and phase. The frequency bins (basis sinusoids) of the DFT can be viewed as narrow-band filters, so the Fourier transform of an input signal can be viewed as passing it through a bank of narrow band-pass filters. This means that on the order of hundreds to thousands of sub-bands are used.

The phase vocoder has found extensive use in computer music composition. Many interesting practical and artistic transformations can be accomplished using the phase vocoder, including nearly artifact-free and independent time and pitch shifting. A technique called *cross synthesis* assumes that one signal is the analysis signal. The time-varying magnitude spectrum of the analysis signal (usually smoothed in the frequency domain) is multiplied by the spectral frames of another input (or filtered) signal (sometimes brightened first by high-frequency emphasis pre-filtering), yielding a composite signal that has the attributes of both. Cross-synthesis has produced the sounds of talking cows, *morphs* between people and cats, trumpet/flute hybrids, etc.

Amazon EX1014 Page 729

17.4 Modal (Damped Sinusoidal) Synthesis

The simplest physical system that does something acoustically (and musically) interesting is the mass/spring/damper (Fig. 17.10 [17.12]. The differential equation describing that system has a solution that is a single exponentially decaying cosine wave. The Helmholtz resonator (large, contained air cavity with a small long-necked opening, like a pop bottle, Fig. 17.11 behaves like a mass/spring/damper system, with the same exponentially damped cosine behavior. The equations describing the behavior of these systems is:

$$d^2y/dt^2 + (r/m)\,dy/dt + (k/m)y = 0\,, \qquad (17.10)$$

$$y(t) = y_0\,e^{(-rt/2m)}\cos\left(t\sqrt{\left[k/m - (r/2m)^2\right]}\right). \qquad (17.11)$$

Of course, most systems that produce sound are more complex than the ideal mass/spring/damper system, or a pop bottle. And of course most sounds are more complex than a simple damped exponential sinusoid. Mathematical expressions of the physical forces (thus the accelerations) can be written for nearly any system, but solving such equations is often difficult or impossible. Some systems have simple enough properties and geometries to allow an exact solution to be written out for their vibrational behavior. An ideal string under tension is one such system.

Here we will resort to some graphical arguments and our prior discussion of the Fourier transform to motivate further the notion of sinusoids in real physical systems. The top of Fig. 17.12 shows a string, lifted from a point in the center (halfway along its length). Below that is shown a set of sinusoidal mode that the center-plucked

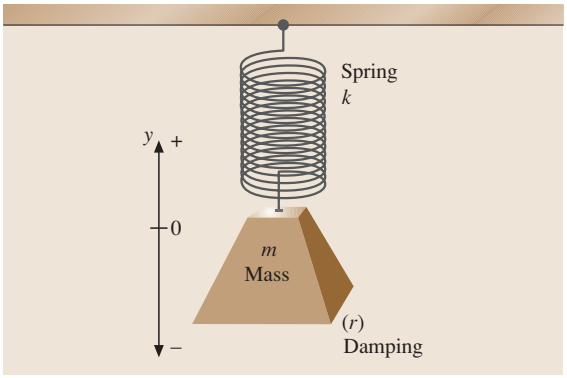


Fig. 17.10 Mass/spring/damper system

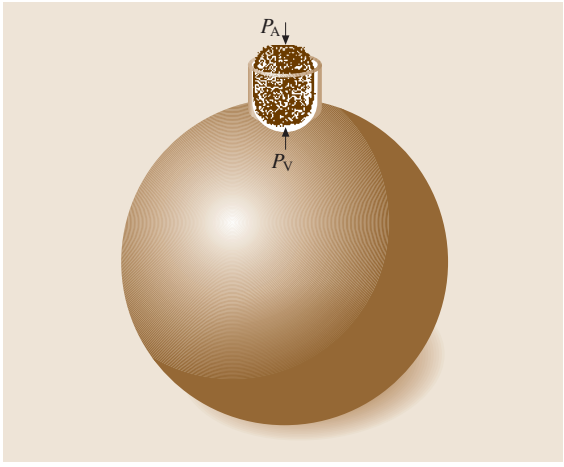


Fig. 17.11 Helmholtz resonator

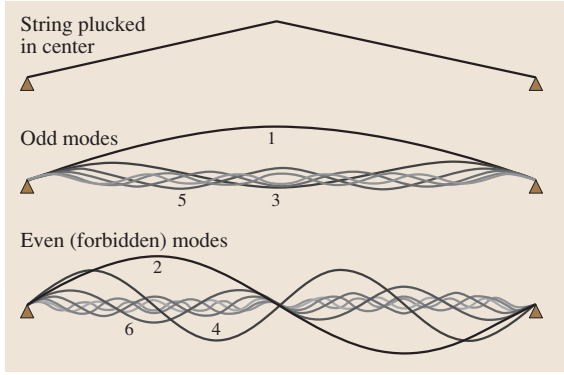


Fig. 17.12 Plucked string (*top*). The *center* shows sinusoidal *modes* of vibration of a center-plucked string. The *bottom* shows the even modes, which would not be excited by the center-plucked condition

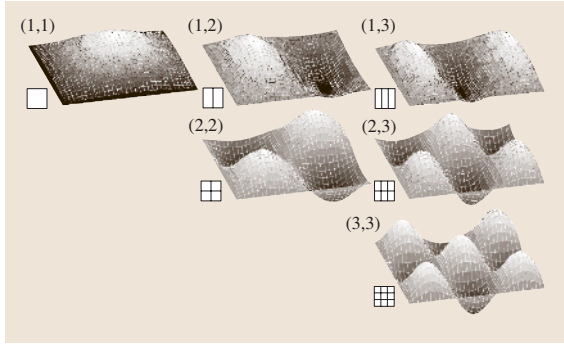


Fig. 17.13 Square-membrane modes

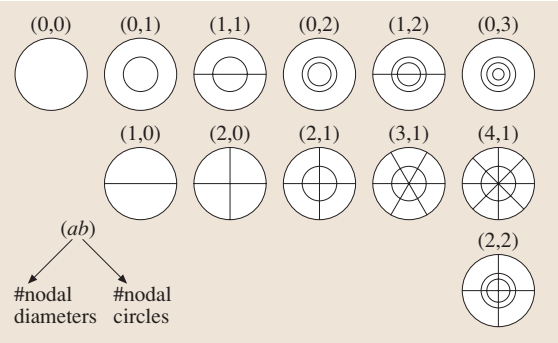


Fig. 17.14 Circular-membrane nodal lines

string vibration would have. These are spatial functions (sine as function of position along the string), but they also correspond to natural frequencies of vibration of the string. At the bottom of Fig. 17.12 is another set of modes that would not be possible with the center-plucked condition, because all of these even modes are restricted to have no vibration in the center of the string, and thus they could not contribute to the triangular shape of the string. These conditions of no displacement, corresponding to the zero crossings of the sine functions, are called *nodes*. Note that the end points are forced nodes of the plucked string system for all possible conditions of excitation.

Physical constraints on a system, such as the pinned ends of a string, and the center plucked initial shape, are known as *boundary conditions*. Spatial sinusoidal solutions like those shown in Fig. 17.12 are called *boundary solutions* (the legal sinusoidal modes of displacement and vibration) [17.13].

Just as one can use Fourier boundary methods to solve the one-dimensional (1-D) string, we can also extend boundary methods to two dimensions. Figures 17.13 and 17.14 show the first few vibration modes of uniform square and circular membranes. The small boxes at the lower left corners of each square modal-shape diagram depict the modes in a purely two-dimensional way, showing lines corresponding to the

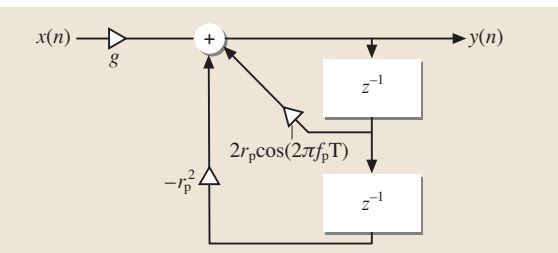


Fig. 17.15 Flexible parametric modal synthesis algorithm

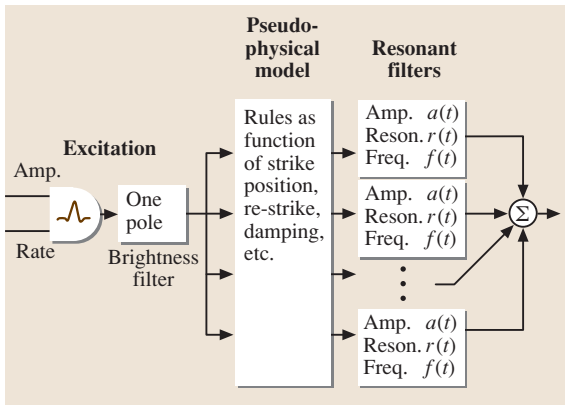


Fig. 17.16 Flexible parametric modal synthesis algorithm

spatial sinusoidal nodes (regions of no displacement vibration). This is how the circular modes are presented. The natural modes must obey the two-dimensional boundary conditions at the edges, but unlike the string, the square membrane modes are not integer-related harmonic frequencies. In fact they obey the relationship

$$f_{mn} = f_{11}\sqrt{\left[(m^2+n^2)/2\right]}\,, \qquad (17.12)$$

where m and n range from 1 to (potentially) infinity, and f_{11} is $c/2L$ (the speed of sound on the membrane divided by the square edge lengths). The circular modes are predictable from mathematics, but have a much more complex form than the square modes.

Unfortunately, circles, rectangles, and other simple geometries turn out to be the only ones for which the boundary conditions yield a closed-form solution in terms of spatial and temporal sinusoidal terms. However, we can measure and model the modes of any system by using the Fourier transform of the sound it produces, and looking for exponentially decaying sinusoidal components.

We can approximate the differential equation describing the mass/spring/damper system of (17.11) by replacing the derivatives (velocity as the derivative of position, and acceleration as the second derivative of position) with sampled time differences (normalized by the sampling interval T seconds). In doing so we would arrive at an equation that is a recursion in past values of $y(n)$, the position variable

$$[y(n)-2y(n-1)+y(n-2)]/T^2 + \frac{r/m[y(n)-y(n-1)]}{T+k/m\,y(n)} = 0\,. \qquad (17.13)$$

Note that if the values of mass, damping, spring constant, and sampling rate are constant, then the coeffi-

Amazon EX1014 Page 731

cients $(2m+Tr)/(m+Tr+T^2k)$ for the single delay, and $m/(m+Tr+T^2k)$ for the twice-delayed signal applied to past y values are constant. Digital signal processing (DSP) engineers would note that a standard infinite impulse response (IIR) recursive filter as shown in Fig. 17.15 can be used to implement (17.13) (the Z^{-1} represents a single sample of delay). In fact, the second order two-pole feedback filter can be used to generate an exponentially decaying sinusoid, called a *phasor* in DSP literature [17.14]. The connection between the second order digital filter and the physical notion of a mode of vibration forms the basis for modal sound synthesis [17.15]. Figure 17.16 shows a general model for modal synthesis of struck/plucked objects, in which an impulsive excitation function is used to excite a number of filters that model the modes. Rules for controlling the modes as a function of strike position, striking object, changes in damping, and other physical constraints are included in the model.

17.5 Subtractive (Source–Filter) Synthesis

Subtractive synthesis uses a complex source wave, such as an impulse, a periodic train of impulses, or white noise, to excite spectral shaping filters. One of the earliest uses of electronic subtractive synthesis dates back to the 1920s and 1930s, with the invention of the *channel vocoder* (or VOiceCODER) [17.16]. In the this device, the spectrum is broken into sections called sub-bands, and the information in each sub-band is converted to a signal representing (generally slowly varying) power. The analyzed parameters are then stored or transmitted (potentially compressed) for reconstruction at another time or physical site. The parametric data representing the information in each sub-band can be manipulated in various ways, yielding transformations such as pitch or time shifting, spectral shaping, cross-synthesis, and other effects. Figure 17.17 shows a block diagram of a channel vocoder. The detected envelopes serve as control signals for a bank of bandpass synthesis filters (identical to the analysis filters used to extract the sub-band envelopes). The synthesis filters have gain inputs that are fed by the analysis control signals.

When used to encode and process speech, the channel vocoder explicitly makes an assumption that the signal being modeled is a single human voice. The *source analysis* block extracts parameters related to finer spectral details, such as whether the sound is pitched (vowel) or noisy (consonant or whispered). If the sound is pitched, the pitch is estimated. The overall energy in the signal is also estimated. These parameters become additional low-bandwidth control signals for the synthesizer. Intelligible speech can be synthesized using only a few hundred numbers per second. An example coding scheme might use eight channel gains + pitch + power, per frame, at 40 frames per second, yielding a total of only 400 numbers per second. The channel vocoder, as designed for speech coding, does not generalize to arbitrary sounds, and fails horribly when the source parameters deviate from expected harmonicity, reasonable pitch range, etc. But the ideas of sub-band decomposition, envelope detection, and driving a synthesis filter bank with control signals give rise to many other interesting applications and implementations of the vocoder concept.

A number of analog hardware devices were produced and sold as musical instrument processing devices in the 1970s and 1980s. These were called simply *vocoders*, but had a different purpose than speech coding. Figure 17.18 shows the block diagram of a cross-

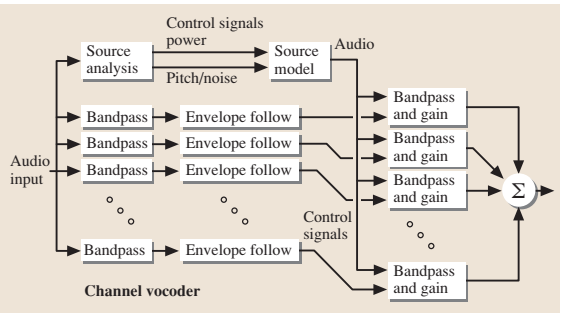


Fig. 17.17 Channel vocoder block diagram

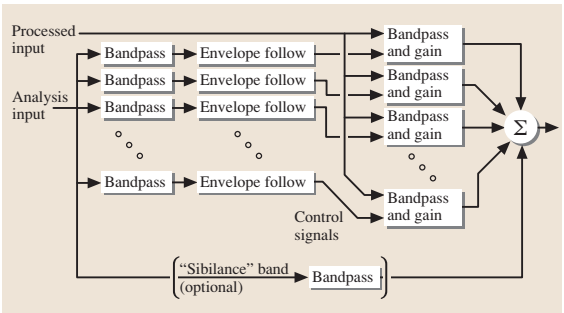


Fig. 17.18 Cross-synthesizing vocoder block diagram

synthesizing vocoder. The main difference is that the parametric source has been replaced by an arbitrary audio signal. The cross-synthesizer can be used to make non-speech sounds talk or sing. A typical example would be feeding a voice into the analysis filter bank, and an electric guitar sound into the synthesis bank audio inputs to make a talking guitar. To be effective, the synthesis audio input should have suitably rich spectral content (for instance, distorted electric guitar works better than undistorted). Some vocoders allowed an additional direct *sibilance band* (like the highest band shown in Fig. 17.18) pass-through of the analyzed sound. This would allow consonants to pass through directly, creating a more intelligible speech-like effect.

Modal synthesis, as discussed before, is a form of subtractive synthesis, but the spectral characteristics of modes are sinusoidal and thus exhibit very narrow spectral peaks. For modeling the gross peaks in a spectrum, which could correspond to weaker resonances, we can exploit the same two-pole resonance filters. This type of source-filter synthesis has been very popular for voice synthesis.

Having its origins and applications in many different disciplines, time-series prediction is the task of estimating future sample values from prior samples. Linear prediction is the task of estimating a future sample (usually the next in the time series) by forming a linear combination of some number of prior samples. Linear predictive coding (LPC) is a technique that can automatically extract the gross spectral features and design filters to match those, and give us a source that we can use to drive the filters [17.17, 18]. Figure 17.19 shows linear prediction in block diagram form. The difference equation for a linear predictor is

$$y(n) = \hat{x}(n+1) = \sum_{i=0}^{m} a_i x(n-i)\,. \tag{17.14}$$

The task of linear prediction is to select the vector of predictor coefficients

$$A = (a_0, a_1, a_2, a_2, \ldots, a_m)$$

such that $\hat{x}(n+1)$ (the estimate) is as close as possible to $x(n+1)$ (the real sample) over a set of samples (often called a frame) $x(0)$ to $x(N-1)$. Usually, "close as possible" is defined by minimizing the mean square error (MSE):

$$MSE = 1/N \sum_{i=0}^{N-1} [\hat{x}(n+1) - x(n+1)]^2\,. \tag{17.15}$$

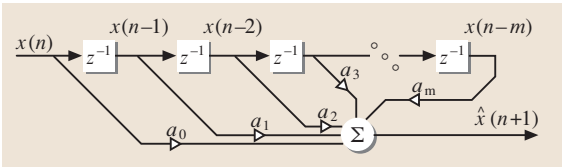


Fig. 17.19 A linear prediction filter

Many methods exist to arrive at the set of predictor coefficients a_i that yield a minimum MSE. The most common uses correlation or covariance data from each frame of samples to be predicted. The *autocorrelation* function, defined as

$$x \diamond x(n) = \sum_{i=0}^{N-1} x(i)x(i+n)\,. \tag{17.16}$$

Autocorrelation computes a time (lag) domain function that expresses the similarity of a signal to delayed versions of itself. It can also be viewed as the convolution of a signal with the time-reversed version of itself. Purely periodic signals exhibit periodic peaks in the autocorrelation function. Autocorrelations of white-noise sequences exhibit only one clear peak at the zero-lag position, and are small (zero for infinite-length sequences) for all other lags. Since the autocorrelation function contains useful information about a signal's similarity to itself at various delays, it can be used to compute the optimal set of predictor coefficients by forming the autocorrelation matrix

$$R = \begin{pmatrix} x\blacklozenge x(0) & x\blacklozenge x(1) & x\blacklozenge x(2) & \ldots & x\blacklozenge x(m) \\ x\blacklozenge x(1) & x\blacklozenge x(0) & x\blacklozenge x(1) & \ldots & x\blacklozenge x(m-1) \\ x\blacklozenge x(2) & x\blacklozenge x(1) & x\blacklozenge x(0) & \ldots & x\blacklozenge x(m-2) \\ x\blacklozenge x(3) & x\blacklozenge x(2) & x\blacklozenge x(1) & \ldots & x\blacklozenge x(m-3) \\ . & . & . & . & \\ x\blacklozenge x(m) & x\blacklozenge x(m-1) & x\blacklozenge x(m-2) & \ldots & x\blacklozenge x(0) \end{pmatrix} \tag{17.17}$$

We can get the least-squares predictor coefficients by forming

$$A = (a_0 a_1 a_2 a_3 \ldots a_m) = PR^{-1}\,, \tag{17.18}$$

where P is the vector of prediction correlation coefficients

$$P = (x\blacklozenge x(1) \;\; x\blacklozenge x(2) \;\; x\blacklozenge x(3) \;\; \ldots \;\; x\blacklozenge x(m+1))\,. \tag{17.19}$$

For low-order LPC (delay order of 6–20 or so), the filter fits itself to the coarse spectral features, and the

Amazon EX1014 Page 733

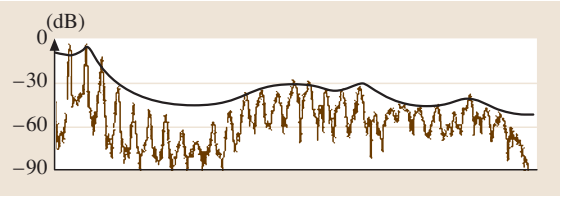


Fig. 17.20 Tenth order LPC filter fit to a voiced 'ooo' spectrum

residue contains the remaining part of the sound that cannot be linearly predicted. A common and popular use of LPC is for speech analysis, synthesis, and compression. The reason for this is that the voice can be viewed as a source-filter model, where a spectrally rich input (pulses from the vocal folds or noise from turbulence) excites a filter (the resonances of the vocal tract). LPC is another form of spectral vocoder (voice coder) as discussed previously, but since LPC filters are not fixed in frequency or shape, fewer bands are needed to model the changing speech spectral shape dynamically.

LPC speech analysis/coding involves processing the signal in blocks and computing a set of filter coefficients for each block. Based on the slowly varying nature of speech sounds (the speech articulators can only move so fast), the coefficients are relatively stable for milliseconds at a time (typically 5–20 ms is used in speech coders). If we store the coefficients and information about the residual signal for each block, we will have captured many of the essential aspects of the signal. Figure 17.20 shows an LPC fit to a speech spectrum. Note that the fit is better at the peak locations than at the valleys. This is due to the least-squares minimization criterion. Missing the mark on low-amplitude parts of the spectrum is not as important as missing it on high-amplitude parts. This is fortunate for audio signal modeling, in that the human auditory system is more sensitive to spectral peaks (formants, poles, resonances) than valleys (zeroes, anti-resonances).

Once we have performed LPC on speech, if we inspect the residual we might note that it is often a stream

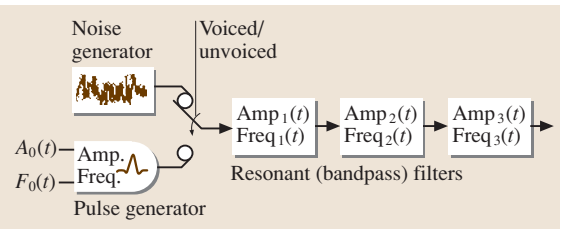


Fig. 17.21 Cascade factored formant subtractive synthesizer

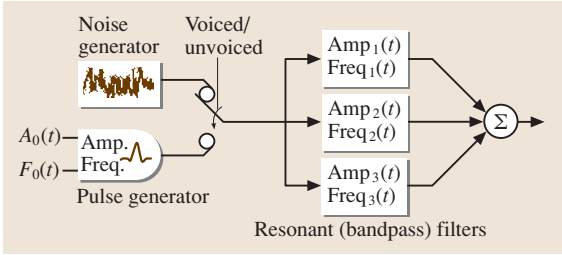


Fig. 17.22 Parallel factored formant subtractive synthesizer

of pulses for voiced speech, or white noise for unvoiced speech. Thus, if we store parameters about the residual, such as whether it is periodic pulses or noise, the frequency of the pulses, and the energy in the residual, then we can get back a signal that is very close to the original. This is the basis of much of modern speech compression. If a signal is entirely predictable using a linear combination of prior samples, and if the predictor filter is doing its job perfectly, we should be able to hook the output back to the input and let the filter predict the rest of the signal automatically. This form of filter, with feedback from output to input, is called *recursive*. The recursive LPC reconstruction is sometimes called *all pole*, to reflect the high-gain poles corresponding to the primary resonances of the vocal tract. The poles do not capture all of the acoustic effects going on in speech, however, such as zeroes that are introduced in nasalization, aerodynamic effects, etc. However, as mentioned before, since our auditory systems are most sensitive to peaks (poles), the LPC representation does quite a good job of capturing the most important aspects of speech spectra.

In recursive resynthesis form, any deviation of the predicted signal from the actual original signal will show up in the error signal, so if we excite the recursive LPC reconstruction filter with the residual signal itself, we can get back the original signal exactly. This is a form of identity analysis/resynthesis, performing deconvolution to separate the source from the filter, and using the residue to excite the filter to arrive at the original signal, by convolution, or using the residue as the input to the filter.

For computer music composition, LPC can be used for cross-synthesis, by keeping the time-varying filter coefficients, and replacing the residual with some other sound source. Using the parametric source model also allows for flexible time and pitch shifting, without modifying the basic timbre. The voiced pulse period can be modified, or the frame rate update of the filter coefficients can be modified, independently. So it is easy to speed up a speech sound while making the pitch lower

Amazon EX1014 Page 734

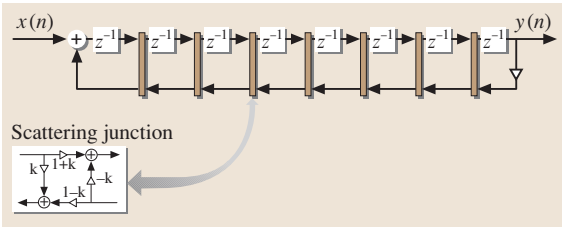


Fig. 17.23 Ladder filter implementation of an all-pole LPC filter

and still retaining the basic spectral shapes of all vowels and consonants.

In decomposing signals into a source and a filter, LPC can be a marvelous aid in analyzing and understanding some sound-producing systems. The recursive LPC reconstruction filter can be implemented in a variety of ways. Three different filter forms are commonly used to perform subtractive voice synthesis [17.19]. The filter can be implemented in series (cascade) as shown in Fig. 17.21, factoring each resonance into a separate filter block with control over center frequency, width, and amplitude. The filter can also be implemented in parallel (separate sub-band sections of the spectrum added together), as shown in Fig. 17.22.

One additional implementation of the resonant filter is the ladder filter structure, which carries with it a notion of one-dimensional spatial propagation as well [17.20]. Figure 17.23 shows a ladder filter realization of a 10th-order IIR filter. We will refer to this type of filter later when we investigate some 1-D waveguide physical models.

17.6 Frequency Modulation (FM) Synthesis

Wave-shaping synthesis involves warping a simple (usually a sine or sawtooth wave) waveform with a nonlinear function or lookup table. One popular form of wave-shaping synthesis, called frequency modulation (FM), uses sine waves for both input and warping waveforms [17.21–23]. Frequency modulation relies on modulating the frequency of a simple periodic waveform with another simple periodic waveform. When the frequency of a sine wave of average frequency f_c (called the carrier), is modulated by another sine wave of frequency f_m (called the modulator), sinusoidal sidebands are created at frequencies equal to the carrier frequency plus and minus integer multiples of the modulator frequency. Figure 17.24 shows a block diagram for simple FM synthesis (one sinusoidal carrier and one sinusoidal modulator). Mathematically, FM is expressed as

$$y(t) = \sin[2\pi t f_c + \Delta f_c \cos(2\pi t f_m)] \,. \qquad (17.20)$$

The index of modulation I is defined as $\Delta f_c / f_c$. The equation as shown is actually phase modulation, but FM and PM can be related by a differentiation/integration of the modulation, and thus do not differ for cosine modulation.

Carson's rule (a rule of thumb) states that the number of significant bands on each side of the carrier frequency (sidebands) is roughly equal to $I + 2$. For example, a carrier sinusoid of frequency 600 Hz, a modulator sinusoid of frequency 100 Hz, and a modulation index of 3 would produce sinusoidal components of frequencies 600, 700, 500, 800, 400, 900, 300, 1000, 200, 1100, and 100 Hz. Inspecting these components reveals that a harmonic spectrum with 11 significant harmonics, based on a fundamental frequency of 100 Hz can be produced by using only two sinusoidal generating functions. Figure 17.25 shows the spectrum of this synthesis.

Selecting carrier and modulator frequencies that are not related by simple integer ratios yields an inharmonic spectrum. For example, a carrier of 500 Hz, modula-

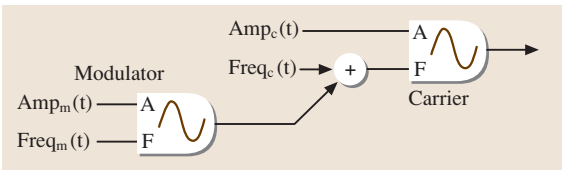


Fig. 17.24 Simple FM (one carrier and one modulator sine wave) synthesis

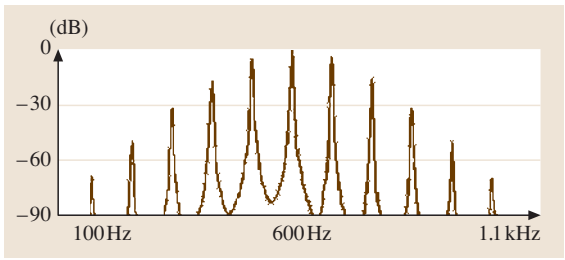


Fig. 17.25 Simple FM with 600 Hz carrier, 100 Hz modulator, and index of modulation of 3

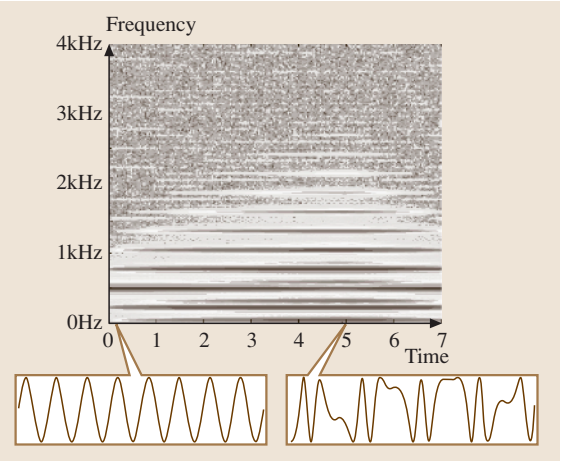


Fig. 17.26 Inharmonic simple FM with 500 Hz carrier and 273 Hz modulator. The index of modulation is ramped from zero to 5 then back to zero

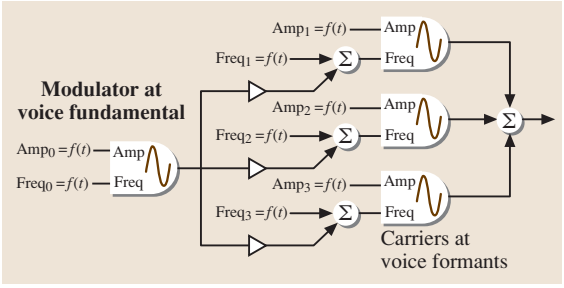


Fig. 17.27 Frequency-modulation voice synthesis diagram

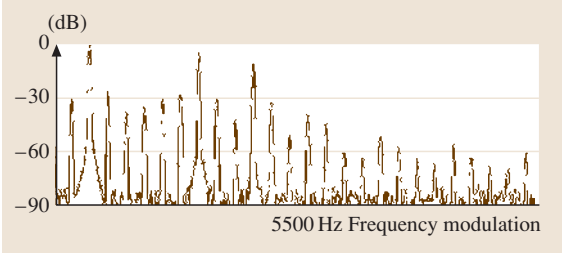


Fig. 17.28 Frequency-modulation voice spectrum

tor of 273 Hz, and an index of 5 yields frequencies of 500 (carrier), 227, 46, 319, 592, 865, 1138, 1411 (negative sidebands), and 773, 1046, 1319, 1592, 1865, 2138, 2411 (positive sidebands). Figure 17.26 shows a spectrogram of this FM tone, as the index of modulation is ramped from zero to 5. The synthesized waveforms at $I = 0$ and $I = 5$ are shown as well.

By setting the modulation index high enough, huge numbers of sidebands are generated, and the aliasing and addition of these (in most cases) results in noise. By careful selection of the component frequencies and index of modulation, and combining multiple carrier/modulator pairs, many spectra can be approximated using FM. However, the amplitudes and phases (described by Bessel functions) of the individual components cannot be independently controlled, so FM is not a truly generic waveform or spectral synthesis method. Using multiple carriers and modulators, connection topologies (algorithms) have been designed for the synthesis of complex sounds such as human voices, violins, brass instruments, percussion, etc. Figures 17.27 and 17.28 show an FM connection topology for human voice synthesis, and a typical resulting spectrum [17.24]. Because of the extreme efficiency of FM, it became popular in the 1980s as a music synthesizer algorithm. Later FM found wide use in early PC sound cards.

17.7 FOFs, Wavelets, and Grains

In a source/filter vocal model such as LPC or parallel/cascade formant subtractive synthesis, periodic impulses are used to excite resonant filters to produce vowels. We could construct a simple alternative model using three, four, or more tables storing the impulse responses of the individual vowel formants. Note that it is not necessary for the tables to contain pure exponentially decaying sinusoids. We could include any aspects of the voice source, etc., as long as those effects are periodic. Formes d'onde formantiques (FOFs, French for formant wave functions) were created for voice synthesis using such tables, overlapped and added at the repetition period of the voice source [17.25]. Figure 17.29 depicts FOF synthesis of a vowel. FOFs are composed of a sinusoid at the formant center frequency, with an amplitude that rises rapidly upon excitation, then decays exponentially. The control parameters define the center frequency and bandwidth of the formant being modeled, and the rate at which the FOFs are generated and added determines the fundamental frequency of the voice.

Note that each individual FOF is a simple wavelet (local and compact wave both in frequency and time).

Amazon EX1014 Page 736

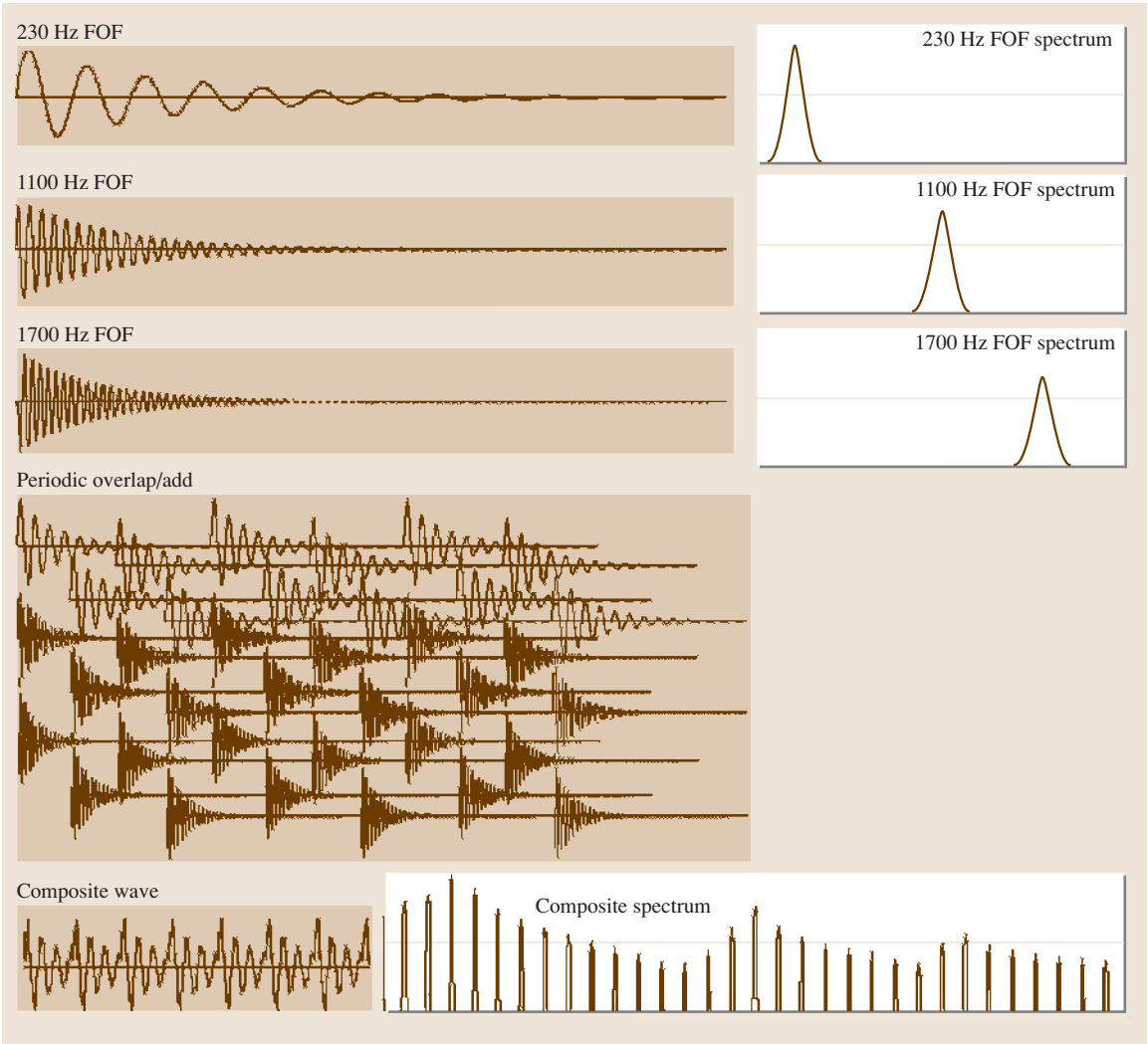


Fig. 17.29 FOF synthesis of a vowel

A family of filter-based frequency transforms known as *wavelet transforms* have been used for the analysis and synthesis of sound. Instead of being based on steady sinusoids like the Fourier transform, wavelet transforms are based on the decomposition of signals into fairly arbitrary wavelets [17.26]. These wavelets can be thought of as mathematical sound grains. Some benefits of wavelet transforms over Fourier transforms are that they can be implemented using fairly arbitrary filter criteria, on a logarithmic frequency scale rather than a linear scale as in the DFT, and that time resolution can be a function of the frequency range of interest. This latter point means that we can say accurate things about high frequencies as well as low. This contrasts to the Fourier transform, which requires the analysis window width to be the same for all frequencies. This means that we must either average out lots of the interesting high-frequency time information in favor of being able to resolve low-frequency information (large window), or opt for good time resolution (small window) at the expense of low-frequency resolution, or perform multiple transforms with different sized windows to catch both time and frequency details. There are a number of fast wavelet transform techniques that allow the subband decomposition to be accomplished in essentially $N \log_2(N)$ time just like the FFT.

Amazon EX1014 Page 737

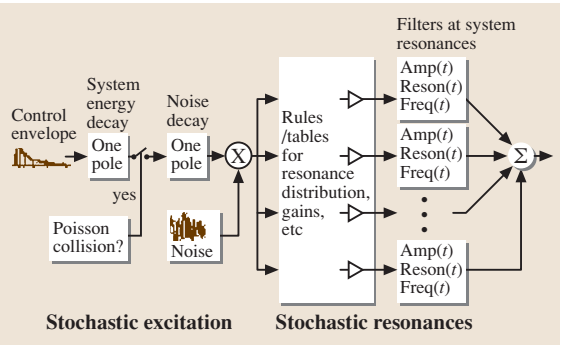


Fig. 17.30 Complete PhISEM model showing stochastic resonances

Most of classical physics can be modeled as objects interacting with each other. Lots of small objects are usually called particles. Granular synthesis involves cutting sound into grains and reassembling them by mixing [17.27]. The grains usually range in length from 5 to 100 ms. The reassembly can be systematic, but often granular synthesis involves randomized grain sizes, locations, and amplitudes. The transformed result usually bears some characteristics of the original sound, just as a mildly blended mixture of fruits still bears some attributes of the original fruits, as well as taking on new attributes due to the mixture. Granular synthesis is mostly used as a compositional and musical-effect type of signal processing, but some also take a physically motivated look at sound grains.

The physically informed stochastic event modeling (PhISEM) algorithm is based on pseudo-random overlapping and adding of parametrically synthesized sound grains [17.28]. At the heart of PhISEM algorithms are particle models, characterized by basic Newtonian equations governing the motion and collisions of point masses as can be found in any introductory physics textbook. By reducing the physical interactions of many particles to their statistical behavior, exhaustive calculation of the position, and velocity of each individual particle can be avoided. By factoring out the resonances of the system, the wavelets can be shortened to impulses or short bursts of exponentially decaying noise. The main PhISEM assumption is that the sound-producing particle collisions follow a Poisson process. Another assumption is that the system energy decays exponentially (the decay of the sound of a maraca after being shaken once, for example). Figure 17.30 shows the PhISEM algorithm block diagram.

The PhISEM maraca synthesis algorithm requires only two random number calculations, two exponential decays, and one resonant filter calculation per sample. Other musical instruments that are quite similar to the maraca include the sekere and cabasa (afuche). Outside the realm of multicultural musical instruments, there are many real-world particle systems that exhibit one or two fixed resonances like the maraca. A bag or box of candy or gum, a salt shaker, a box of wooden matches, and gravel or leaves under walking feet all fit well within this modeling technique.

In contrast to the maraca and guiro-like gourd resonator instruments, which exhibit one or two weak resonances, instruments such as the tamborine (timbrel) and sleigh bells use metal cymbals, coins, or bells suspended on a frame or stick. The interactions of the metal objects produce much more pronounced resonances than the maraca-type instruments, but the Poisson event and exponential system energy statistics are similar enough to justify the use of the PhISEM algorithm for synthesis. To implement these in PhISEM, more filters are used to model the individual partials, and at each collision, the resonant frequencies of the filters are randomly set to frequencies around the main resonances. Other sounds that can be modeled using stochastic filter resonances include bamboo wind chimes (related to a musical instrument as well in the Javanese anklung) [17.29].

17.8 Physical Modeling (The Wave Equation)

One way to model a vibrating string physically would be to build it up as a chain of coupled masses and springs, as shown in Fig. 17.31. The masses slide on frictionless guide rods to restrict their motion to transverse (at right angles to the length), as is the case in the ideal string. In fact, some synthesis projects have used mass/spring/damper systems to represent arbitrary objects [17.30].

If we force the masses to move only up and down, and restrict them to small displacements, we do not actually have to solve the acceleration of each mass as a function of the spring forces pushing them up and down. There is a simple differential equation that completely describes the motions of the ideal string:

$$d^2y/dx^2 = (1/c^2)d^2y/dt^2 \,. \qquad (17.21)$$

Amazon EX1014 Page 738

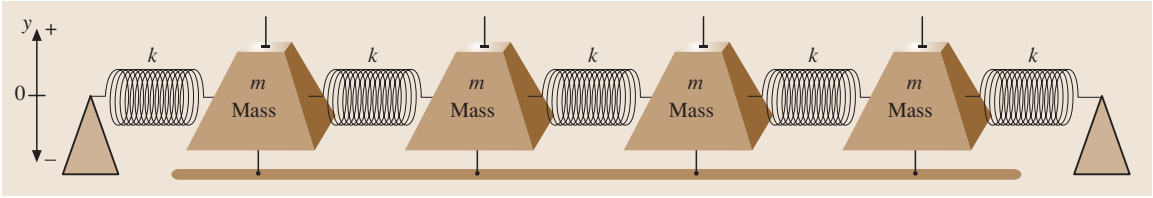


Fig. 17.31 Mass/spring string network, with frictionless *guide rods* to restrict motion to one dimension

This equation (called the wave equation) means that the acceleration (up and down) of any point on the string is equal to a constant times the curvature of the string at that point. The constant c is the speed of wave motion on the string, and is proportional to the square root of the string tension, and inversely proportional to the square root of the mass per unit length. This equation could be solved numerically, by sampling it in both time and space, and using the difference approximations for acceleration and curvature (much like we did with the mass/spring/damper system earlier). With boundary conditions (like rigid terminations at each end), we could express the solution of this equation as a Fourier series, as we did earlier in graphical form (Fig. 17.12). However, there is one more wonderfully simple solution to (17.21), given by

$$y(x,t) = y_L(t + x/c) + y_r(t - x/c)\,. \qquad (17.22)$$

This equation says that any vibration of the string can be expressed as a combination of two separate traveling waves, one traveling left (y_L) and one traveling right (y_r). They move at the rate c, which is the speed of sound propagation on the string. For an ideal (no damping) string and ideally rigid boundaries at the ends, the wave reflects with an inversion at each end, and would travel back and forth indefinitely. This view of two traveling waves summing to make a displacement wave gives rise to the waveguide filter technique of modeling the vibrating string [17.31, 32]. Figure 17.32 shows a waveguide filter model of the ideal string. The two delay lines model the propagation of left- and right-going traveling waves. The conditions at the ends model the reflection of the traveling waves at the ends. The −1 in the left reflection models the reflection with inversion of a displacement wave when it hits an ideally rigid termination (like a fret on a guitar neck). The −0.99 on the right reflection models the slight loss that happens when the wave hits a termination that yields slightly (like the bridge of the guitar which couples the string motion to the body), and models all other losses the wave might experience (internal damping in the string, viscous losses as the string cuts the air, etc.) in making its round-trip path around the string.

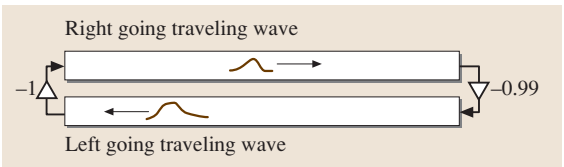


Fig. 17.32 Waveguide string modeled as two delay lines

Figure 17.33 shows the waveguide string as a digital filter block diagram. The $Z^{-P/2}$ blocks represent a delay equal to the time required for a wave to travel down the string. Thus a wave completes a round trip each P samples (down and back), which is the fundamental period of oscillation of the string, expressed in samples. Initial conditions can be injected into the string via the input $x(n)$. The output $y(n)$ would yield the right-going traveling wave component. Of course, neither of these conditions is actually physical in terms of the way a real

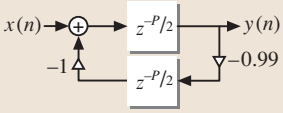


Fig. 17.33 Filter view of a waveguide string

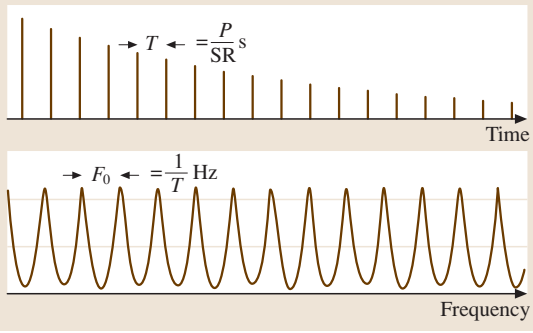


Fig. 17.34 Impulse response and spectrum of a comb (string) filter

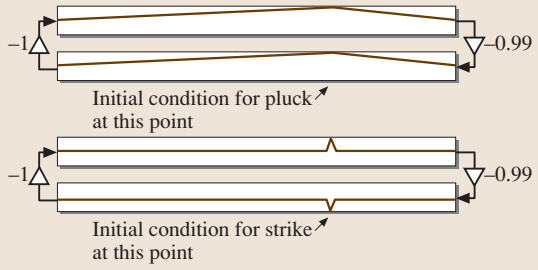


Fig. 17.35 Waveguide pluck and strike initial conditions

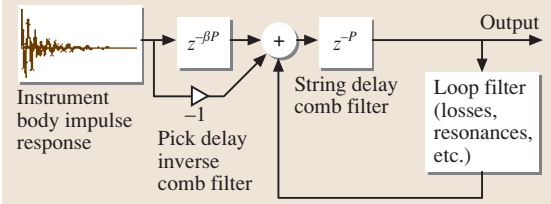


Fig. 17.37 Commuted plucked string model

string is plucked and listened to, but feeding the correct signal into x is identical to loading the delay lines with a predetermined shape.

The impulse response and spectrum of the filter shown in Fig. 17.32 is shown in Fig. 17.34. As would be expected, the impulse response is an exponentially decaying train of pulses spaced $T = P/\text{SRATE}$ seconds apart, and the spectrum is a set of harmonics spaced $F_0 = 1/T$ Hz apart. This type of filter response and spectrum is called a *comb filter*, so named because of the comb-like appearance of the time-domain impulse response, and of the frequency-domain harmonics.

The two delay lines taken together are called a *waveguide filter*. The sum of the contents of the two delay lines is the displacement of the string, and the difference of the contents of the two delay lines is the velocity of the string. If we wish to pluck the string, we simply need to load 1/2 of the initial-shape magnitude into each of the upper and lower delay lines. If we wish to strike the string, we would load in an initial velocity by entering a positive pulse into one delay line and a negative pulse into the other (difference = initial velocity, sum = initial position = 0). These conditions are shown in Fig. 17.35.

Figure 17.36 shows a relatively complete model of a plucked string simulation system using digital filters. The inverse comb filters model the nodal effects of picking, and the output of an electrical pickup, emphasizing certain harmonics and forbidding others based on the pick (pickup) position [17.33]. Output channels for the pickup position and body radiation are provided separately. A solid-body electric guitar would have no direct radiation and only pickup output(s), while a purely acoustic guitar would have no pickup output, but possibly a family of directional filters to model body radiation in different directions [17.34].

We should note that everything in the block diagram of Fig. 17.35 is linear and time invariant (LTI). Linearity means that for input x and output y, if $x_1(t) \rightarrow y_1(t)$ and $x_1(t) \rightarrow y_2(t)$, then $x_2(t)+x_2(t) \rightarrow y_1(t)+y_2(t)$, for any functions $x_{1,2}$ (signals add and scale linearly). Time invariant means that if $x(t) \rightarrow y(t)$, then $x(t-T) \rightarrow y(t-T)$ for any $x(t)$ and any T (the system doesn't change with time). Given that the LTI condition is satisfied, we can then happily (legally) commute (swap the order) of any of the blocks. For example, we could put the pick position filter at the end, or move the body radiation filter to the beginning. Since the body radiation filter is just the impulse response of the instrument body, we can actually record the impulse response with the strings damped, and use that as the input to the string model. This is called *commuted synthesis* [17.35]. Figure 17.37 shows a commuted synthesis plucked string model.

Adding a model of bowing friction allows the string model to be used for the violin and other bowed strings. This focused nonlinearity is what is responsible for turning the steady linear motion of a bow into an oscillation of the string [17.36, 37]. The bow sticks to the string

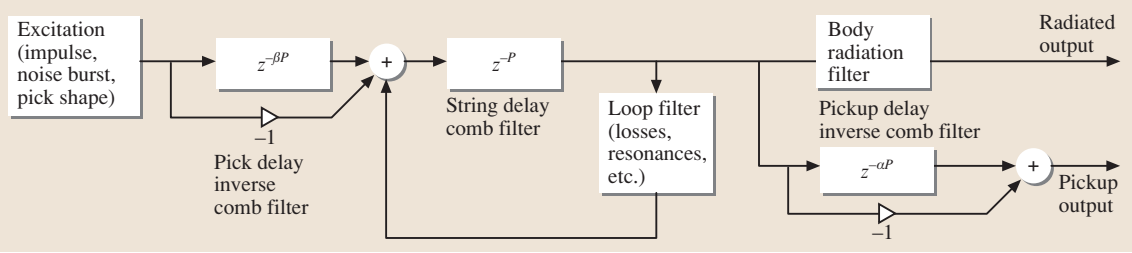


Fig. 17.36 Fairly complete digital filter simulation of the plucked string system

Amazon EX1014 Page 740

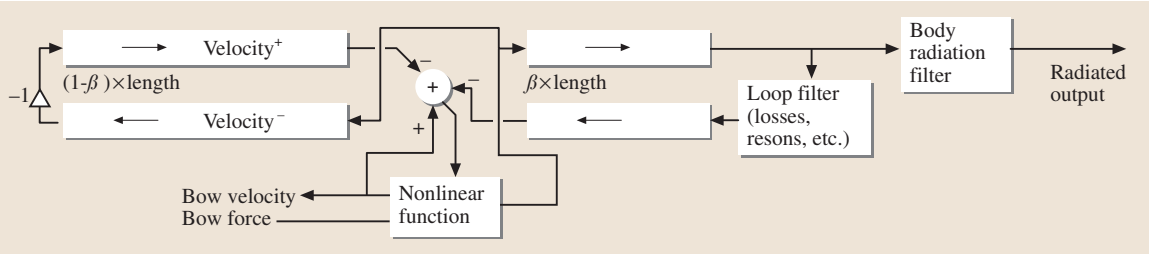


Fig. 17.38 Bowed string model

for a while, pulling it along, then the forces become too great and the string breaks away, flying back toward rest position. This process repeats, yielding a periodic oscillation. Figure 17.38 shows a simple bowed string model, in which string velocity is compared to bow velocity, then put through a nonlinear friction function controlled by bow force. The output of the nonlinear function is the velocity input back into the string.

In mathematically describing the air within a cylindrical acoustic tube (like a trombone slide, clarinet bore, or human vocal tract), the defining equation is:

$$\mathrm{d}^2P/\mathrm{d}x^2 = \left(1/c^2\right)\left(\mathrm{d}^2P/\mathrm{d}t^2\right) \qquad (17.23)$$

which we note has exactly the same form as (17.21), except that the displacement y is replaced by the pressure P. A very important paper in the history of physical modeling [17.36] noted that many musical instruments can be characterized as a linear resonator, modeled by filters such as all-pole resonators or waveguides, and a single nonlinear oscillator like the reed of the clarinet, the lips of the brass player, the jet of the flute, or the bow–string friction of the violin. Since the wave equation says that we can model a simple tube as a pair of bidirectional delay lines (waveguides), then we can build models using this simple structure. If we would like to do something interesting with a tube, we could use it to build a flute or clarinet. Our simple clarinet model might look like the block diagram shown in Fig. 17.39.

To model the reed, we assume that the mass of the reed is so small that the only thing that must be considered is the instantaneous force on the reed (spring). The pressure inside the bore P_b is the calculated pressure in our waveguide model, the mouth pressure P_m is an external control parameter representing the breath pressure inside the mouth of the player (Fig. 17.40). The net force acting on the reed/spring can be calculated as the difference between the internal and external pressures, multiplied by the area of the reed (pressure is force per unit area). This can be used to calculate a reed opening position from the spring constant of the reed. From the reed opening, we can compute the amount of pressure that is allowed to leak into the bore from the player's mouth. If bore pressure is much greater than mouth pressure, the reed opens far. If the mouth pressure is much greater than the bore pressure, the reed slams shut. These two extreme conditions represent an asym-

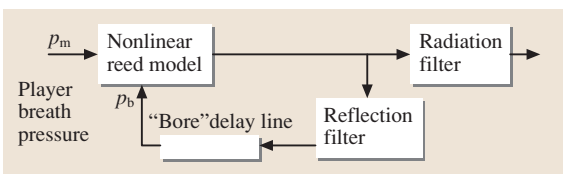


Fig. 17.39 Simple clarinet model

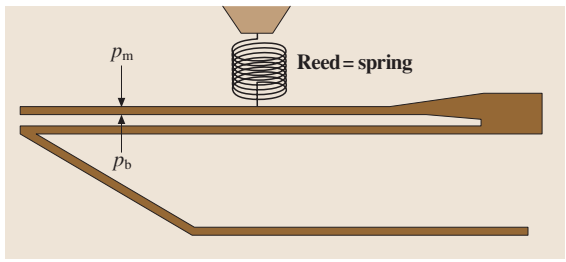


Fig. 17.40 Reed model

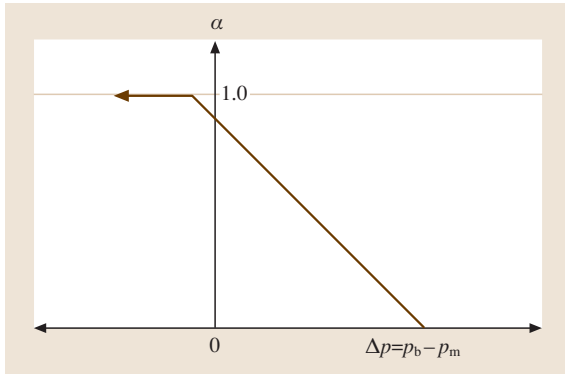


Fig. 17.41 Reed reflection table

Amazon EX1014 Page 741

metric nonlinearity in the reed response. Even a grossly simplified model of this nonlinear spring action results in a pretty good model of a clarinet [17.37]. Figure 17.41 shows a plot of a simple reed reflection function (as seen from within the bore) as a function of differential pressure. Once this nonlinear signal-dependent reflection coefficient is calculated (or looked up in a table), the right-going pressure injected into the bore can be calculated as $P_b^+ = \alpha P_b^- + (1-\alpha)P_m$.

The clarinet is open at the bell end, and essentially closed at the reed end. This results in a reflection with inversion at the bell and a reflection without inversion (plus any added pressure from the mouth through the reed opening) at the reed end. These odd boundary conditions cause odd harmonics to dominate in the clarinet spectrum.

We noted that the ideal string equation and the ideal acoustic tube equation are essentially identical. Just as there are many refinements possible to the plucked string model to make it more realistic, there are many possible improvements for the clarinet model. Replacing the simple reed model with a variable mass/spring/damper allows the modeling of a lip reed as is found in brass instruments. Replacing the reed model with an air-jet model allows the modeling of flute and recorder-like instruments. With all wind or friction (bowed) excited resonant systems, adding a little noise in the reed/jet/bow region adds greatly to the quality (and behavior) of the synthesized sound.

In an ideal string or membrane, the only restoring force is assumed to be the tension under which they are stretched. We can further refine solid systems such as strings and membranes to model more-rigid objects, such as bars and plates, by noting that the more-rigid objects exhibit internal restoring forces due to their stiffness. We know that if we bend a stiff string, it wants to return back to straightness even when there is no tension on the string. Cloth string or thread has almost no stiffness. Nylon and gut strings have some stiffness, but not as much as steel strings. Larger-diameter strings have more stiffness than thin ones. In the musical world, piano strings exhibit the most stiffness. Stiffness results in the restoring force being higher (thus the speed of sound propagation as well) for high frequencies than for low. So the traveling wave solution is still true in stiff systems, but a frequency-dependent propagation speed is needed

$$y(x,t) = y_l\left[t + x/c(f)\right] + y_r\left[t - x/c(f)\right] . \quad (17.24)$$

and the waveguide filter must be modified to simulate frequency-dependent delay, as shown in Fig. 17.42.

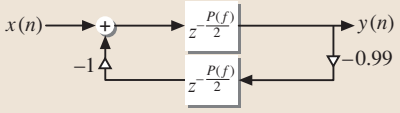


Fig. 17.42 Stiffness-modified waveguide string filter

For basic stiff strings, a function that predicts the frequencies of the partials has this form:

$$f_n = n\left[f_0\left(1 + Bn^2\right)\right] , \quad (17.25)$$

where B is a number slightly greater than 0, equal to zero for perfect harmonicity (no stiffness), and increasing for increasing stiffness. Typical values of B are 0.00001 for guitar strings, and 0.004 or so for piano strings. This means that $P(f)$ should modify P by the inverse of the $(1 + Bn^2)^{1/2}$ factor.

Unfortunately, implementing the $Z^{-P(f)/2}$ frequency-dependent delay function is not simple, especially for arbitrary functions of frequency. One way to implement the $P(f)$ function is by replacing each of the Z^{-1} with a first-order all-pass (phase) filter, as shown in Fig. 17.43 [17.33]. The first-order all-pass filter has one pole and one zero, controlled by the same coefficient. The all-pass implements a frequency-dependent phase delay, but exhibits a gain of 1.0 for all frequencies. The coefficient α can take on values between -1.0 and 1.0. For $\alpha = 0$, the filter behaves as a standard unit delay. For $\alpha > 0.0$, the filter exhibits delays longer than one sample, increasingly long for higher frequencies. For $\alpha < 0.0$ the filter exhibits delays shorter than one sample, decreasingly so for high frequencies.

It is much less efficient to implement a chain of all-pass filters than a simple delay line. But for weak stiffness it is possible that only a few all-pass sections will provide a good frequency-dependent delay. Another option is to implement a higher-order all-pass filter, designed to give the correct stretching of the upper frequencies, added to simple delay lines to give the correct longest bulk delay required.

For very stiff systems such as rigid bars, a single waveguide with all-pass filters is not adequate to give enough delay, or far too inefficient to calculate. A technique called *banded waveguides* employs

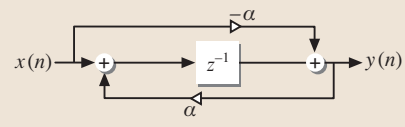


Fig. 17.43 First-order all-pass filter

Amazon EX1014 Page 742

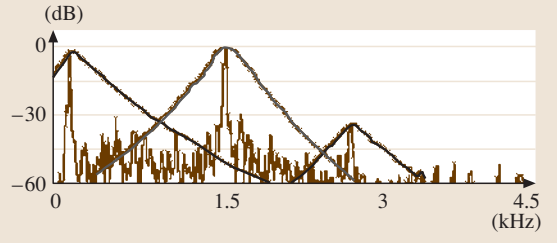


Fig. 17.44 Banded decomposition of a struck-bar spectrum

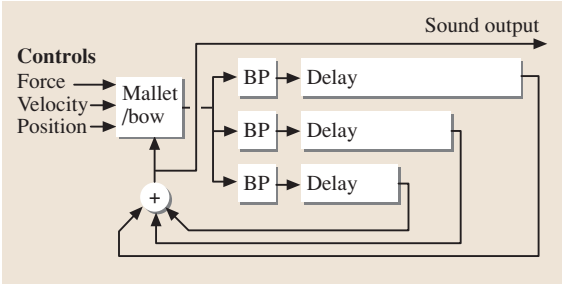


Fig. 17.45 Banded waveguide model

sampling in time, space, and frequency to model stiff one-dimensional systems [17.38]. This can be viewed as a hybrid of modal and waveguide synthesis, in that each waveguide models the speed of sound in the region around each significant mode of the system. As an example, Fig. 17.44 shows the spectrum of a struck marimba bar, with additional bandpass filters superimposed on the spectrum, centered at the three main modes. In the banded waveguide technique, each mode is modeled by a bandpass filter, plus a delay line to impose the correct round-trip delay, as shown in Fig. 17.45.

17.9 Music Description and Control

The musical instrument digital interface (MIDI) standard, adopted in 1984, revolutionized electronic music, and also profoundly affected the computer industry [17.39]. A simple two-wire serial electrical connection standard allows interconnection of musical devices and computers over cable distances of up to 15 m (longer over networks and extensions to the basic MIDI standard). The MIDI software protocol is best described as musical keyboard gestural, meaning that the messages carried over MIDI are the gestures that a pianist or organist uses to control a traditional keyboard instrument. There is no time information contained in the basic MIDI messages; they are intended to take effect as soon as they come over the wire. As such it is a real-time gestural protocol, and can be adapted for real-time non-musical sound synthesis applications. There are complaints about MIDI, however, mostly related to limited control bandwidth (approximately 1000–1500 continuous messages per second maximum), and the keyboard-centric bias of the messages.

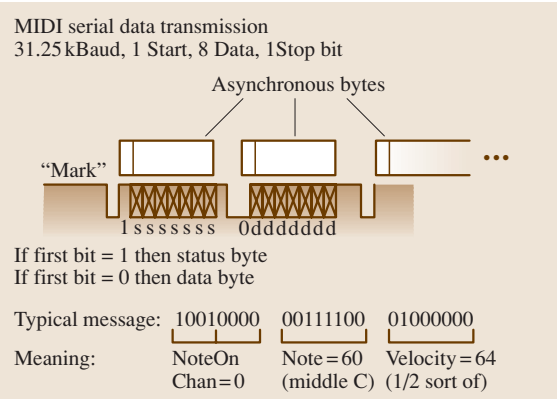


Fig. 17.46 MIDI software transmission protocol

Basic MIDI message types include NoteOn and NoteOff, sustain pedal up and down, amount of modulation (vibrato), pitch bend, key pressure (also called AfterTouch), breath pressure, volume, pan, balance, reverberation amount, and others. NoteOn and NoteOff messages carry a note number corresponding to a particular piano key, and a velocity corresponding to how hard that key is hit. Figure 17.46 shows the software serial format of MIDI, and gives an example of a NoteOn message.

Another MIDI message is program change, which is used to select the particular sound being controlled in the synthesizer. MIDI provides for 16 channels, and the channel number is encoded into each message. Each channel is capable of one type of sound, but possibly many simultaneous *voices* (a voice is an individual sound) of that same sound. Channels allow instruments and devices to all listen on the same network, and choose to respond to the messages sent on particular channels.

General MIDI (1991), and the standard MIDI file specifications serve to extend MIDI and made it even more popular [17.40]. General MIDI helps to assure

Amazon EX1014 Page 743

the performance of MIDI on different synthesizers, by specifying that a particular program (the algorithm for producing the sound of an instrument) number must call up a program the approximates the same instrument sound on all general MIDI-compliant synthesizers. There are 128 such defined instrument sounds available on MIDI channels 1–9 and 11–16. For example, MIDI program 1 is grand piano, and 57 is trumpet. Some of the basic instrument sounds are sound effects. Program #125 is a telephone ring, #126 is helicopter, #127 is applause, and #128 is a gunshot sound. On general MIDI channel 10, each note is mapped to a different percussion sound. For example, the bass drum is note number 35, and the cowbell is note number 56.

The MIDI file formats provide a means for the standardized exchange of musical/sonic information. A MIDI level 0 file carries the basic information that would be carried over a MIDI serial connection, which is the basic bytes in order, with added time stamps. Time stamps allow a very simple program to play back MIDI files. A MIDI level 1 file is more suited to manipulation by a notation program or MIDI sequencer (a form of multi-track recorder program that records and manipulates MIDI events). Data is arranged by *tracks*, (each track transmits on only one MIDI channel) which are the individual instruments in the virtual synthesized orchestra. Meta messages allow for information which is not actually required for a real-time MIDI playback, but might carry information related to score markings, lyrics, composer and copyright information, etc.

Recognizing the limitations of fixed PCM samples in music synthesizers, the downloadable sounds (DLS) specification was added to MIDI in 1999. This provides a means to load PCM into a synthesizer, then use MIDI commands to play it back. Essentially any instrument/percussion sound can be replaced with arbitrary PCM, making it possible to control large amounts of recorded sounds. The emergence of software synthesizers, on PC platforms has made the use of DLS more feasible than in earlier days when the samples were contained in a static ROM within a hardware synthesizer.

The simplicity and ubiquity of MIDI, with thousands of supporting devices and software systems, leads many to use it for all sorts of projects and systems. While not being directly compatible with popular music synthesizers, using MIDI for the control of synthesis still allows the use of the large variety of MIDI control signal generators and processors, all of which emit standard and user-programmable MIDI messages. Examples include MIDI sequencers and other programs for general-purpose computers, as well as breath controllers, drum controllers, and fader boxes with rows of sliders and buttons.

Sadly, one thing that is still missing from MIDI at the time of writing is the ability to download arbitrary synthesis algorithms. Beyond MIDI, there are a number of emerging standards for sound and multimedia control. As mentioned in the last section, one of the things MIDI lacks is high-bandwidth message capability for handling many streams of gestural/model control data. Another complaint about MIDI relates to the small word sizes, such as semitone quantized pitches (regular musical notes), and only 128 levels of velocity, volume, pan, and other controllers. MIDI does support some extended precision messages, allowing 14 bits rather than 7. Other specifications have proposed floating-point representations, or at least larger integer word sizes. A number of new specifications and systems have arisen recently.

One system for describing sounds themselves is called the sound description interchange format (SDIF 1999) [17.41], originally called the spectral description interchange file (SDIF) format. It is largely intended to encapsulate the parameters arising from spectral analysis, such as spectral modeling sinusoidal/noise tracks, or source/filter model parameters using LPC.

Related to SDIF, but on the control side, is open sound control (OSC, 2002) [17.41]. Open sound control allows for networked, extended-precision, object-oriented control of synthesis and processing algorithms over high-bandwidth transmission control protocol/internet protocol (TCP/IP), FireWire, and other high-bandwidth protocols.

As part of the moving-picture experts group 4 (MPEG4) standard, the structured audio orchestra language (SAOL, 1999) allows for parametric descriptions of sound algorithms, controls, effects processing, mixing, and essentially all layers of the audio synthesis and processing chain [17.42]. One fundamental notion of structured audio is that the more understanding we have about how a sound is made, the more flexible we are in compressing, manipulating, and controlling it.

Amazon EX1014 Page 744

17.10 Composition

The history of computer languages for music dates back well into the 1950s. Beginning with Max Mathew's Music I, many features of these languages have remained constant, such as the notion of unit generators as discussed before, and the idea of instruments within an orchestra, controlled by a score. Early popular languages include Music V (a direct descendent of Music I) [17.5], Csound, CMix, and CMusic (written in C) [17.43]. More recent languages developed in the last decade include SuperCollider [17.44], Jsyn (in Java) [17.45], STK in C++ [17.46], Aura [17.47], Nyquist [17.48], and ChucK [17.49].

One class of languages that has advanced education and participation in computer music are graphical systems such as MAX/MSP [17.50] and Pure Data (PD) [17.51]. These software systems allow novice (and expert) users to construct patches by dragging unit generator icons around on the desktop, connecting them with virtual patch cords (Fig. 17.47).

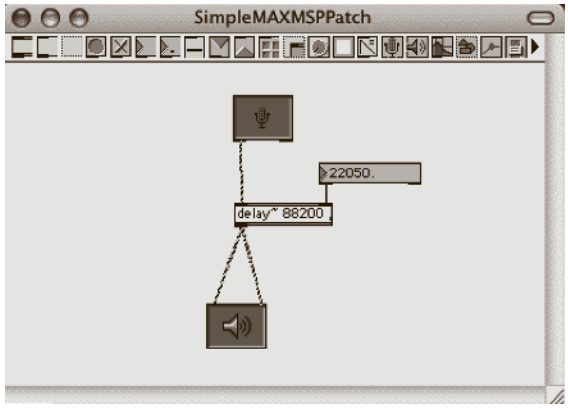


Fig. 17.47 A simple MAX/MSP patch connects a microphone through a 1/2 s (22050 samples at 44.1 k sample rate) delay to the speaker output

17.11 Controllers and Performance Systems

Music performance is structured real-time control of sound, to achieve an aesthetic goal (hopefully for both the performers and audience). So one obvious application area for real-time computer sound synthesis and processing is in creating new forms of musical expression. Exciting developments in the last few years have emerged in the area of new controllers and systems for using parametric sound in real-time music performance, or in interactive installation art.

Figure 17.48 shows three recent real-time music performance systems, all using parametric digital sound synthesis and a variety of sensors mounted on the instruments and players. The author's DigitalDoo is a sensor/speaker enhanced/augmented didgeridoo. Dan Trueman's bowed sensor speaker array (BoSSA) resulted from work in modeling, refining, and redefining the violin player's interface [17.52]. Curtis Bahn's 'r!g' consists of a sensor enhanced upright (stick) bass, pedal boards, and other controllers. All three of these new instruments use spherical speakers to allow flexible control of the sound radiation into the performance space [17.53]. These instruments have been used with computers to control percussion sounds, animal noises, string models, flute models, and a large va-



Fig. 17.48 Cook's DigitalDoo, Trueman's BoSSA, Bahn's 'r!g'



Fig. 17.49 Gigapop networked concert between Princeton (shown on screen to *left*) and Montreal (players on stage). The right screen shows real-time graphics generated from gestural data captured from sensors on the instruments

riety of other sample-based and parametric synthesis algorithms.

Various conducting systems, including Max Mathew's radio baton (Stanford's Center for Computer Research in Music and Acoustics) [17.54], and Theresa Marrin Nakra's conducting jacket (MIT Media Lab) [17.55] allow novices and experts to conduct music performed by the computer. Also, the growth and development of computer music has helped to advance greatly the technological capabilities of the familiar karaoke industry. Some modern karaoke systems now can deduce the song that is being sung and retrieve the appropriate accompaniment within just a few notes. Also, these systems can automatically correct the pitch and vocal qualities of a bad singer, making them sound more expert and professional. The next section will describe some of the algorithms being applied now in karaoke and other automatic accompaniment systems.

Some researchers and performers feel that there is significant potential in networked computer music performances and interactions. Projects range from MIT's Brain Opera [17.56], which allowed observers to participate and contribute content to the performance via the web, through the Auricle website, which is a kind of audio analysis/synthesis enhanced chat room [17.57]. Stanford's SoundWire Project [17.58] and Princeton's GigaPop Project [17.59] (Fig. 17.49) are aimed at synchronized multi-site live performance through high-bandwidth low-latency streaming of uncompressed audio and video.

17.12 Music Understanding and Modeling by Computer

Over the years, new research and applications have arisen in the areas of content-based audio analysis, retrieval, and modeling. Also called music information retrieval (MIR), machine listening, and other terms, the main tasks involve segmenting, separating, classifying, and modeling the textures and structures present in pieces of music or audio.

Audio segmentation is the process of breaking up an audio stream into sections that are perceptually different from adjacent sections. The audio texture within a given segment is relatively stable. Examples of segment boundaries could be a transition from background sound texture to a human beginning to speak over that background. Another segment boundary might occur when the scene changes, such as leaving an office building lobby and going outside onto a busy street. In movie production, the audio scene often changes just prior to the visual scene, allowing a noncausal mode of detecting a visual scene change.

Segmentation [17.60] can be accomplished in two primary ways: blind segmentation based on sudden changes in extracted audio features, and classification-based segmentation based on comparing audio features to a set of trained target feature values. The blind method works well and is preferred when the segment textures are varied and unpredictable, but requires the setting of thresholds to yield the best results. The classification-based method works well on a corpus of pre-labeled data, such as speech/music, musical genres, indoor/outdoor scenes, etc. databases. Both methods require the extraction of audio features.

Audio feature extraction is the process of computing a compact numerical representation of a sound segment. A variety of audio features exist, and have been used in systems for speech recognition, music/speech discrimination, musical genre (rock, pop, country, classical, jazz, etc.) labeling, and other audio classification tasks. Most features are extracted from short moving windows (5–100 ms in length, moving along at a rate of 5–20 windows per second) by using the short-time Fourier transform, wavelets, or compressed data coefficients (such as MP3). Each feature described below can be computed at different time resolutions, and the value of each feature, along with the mean

and variance of the features can be used as features themselves.

- One common feature for segmentation is the gross power in each window. If the audio stream suddenly gets louder or softer, then there is a high likelihood that something different is occurring. In speech recognition and some other tasks, however, we would like the classification to be loudness invariant (over some threshold used to determine if anyone is speaking).
- Another feature is the spectral centroid, which is closely related to the brightness of the sound, or the relative amounts of high- and low-frequency energy. The spectral roll-off is another important feature that captures more information about the brightness of an audio signal.
- Spectral flux is the amount of frame-to-frame variance in the spectral shape. A steady sound or texture will exhibit little spectral flux, while a modulating sound will exhibit more flux.
- One popular set of features used for speech and speaker recognition are mel-frequency cepstral coefficients, which are a compact (between 4 and 10 numbers) representation of spectral shape.
- For multi-time-scale analysis systems, the low-energy feature is often used. This feature is defined as the percentage of small analysis windows that contain less power than the average over the larger window that includes the smaller windows. This is a coarser-time-scale version of flux, but computed only for energy.
- One other commonly used feature is zero-crossing rate, which is a simple measure of high-frequency energy.
- For music query/recognition, features such as the parametric pitch histogram, and beat/periodicity histogram can be calculated and used.

Selection of the correct feature set for a given task has proven to be an important part of building successful systems for machine audio understanding. For a fixed corpus, computing many features (40 dimensions or more), then using principal-components analysis (reduction of dimensionality through regression) has proven successful for reducing the dimensionality of the feature/search space. As an example, mapping the first three principal components onto color (red, green, blue) allows the automatic coloration of sound plots, as shown in Fig. 17.50. The three *timbregrams* on the left are speech, and the three on the right are classical music. The top two right timbergrams are orchestral recordings, while the lower right one is opera singing (hence the color similarity to speech).

One holy grail in computer music research is automatic transcription, where in the ideal case a computer could generate a perfect musical score by analyzing only a recording of the music. This problem has been around for many decades, and has generated significant funding for computer music researchers. While research still continues on this difficult and unsolved problem, the side effects of the research have given us a large number of developments in signal

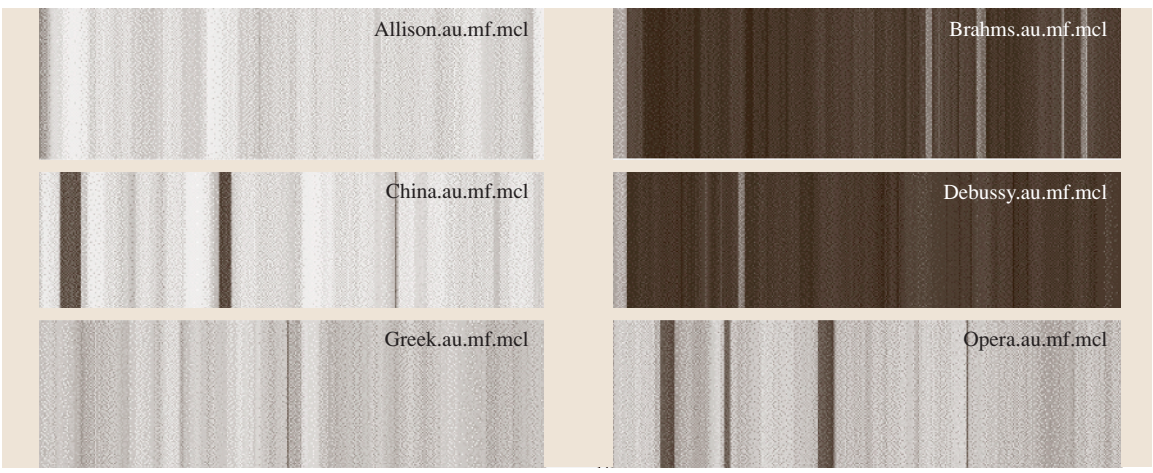


Fig. 17.50 *Timbregrams* of speech (*left*) and music (*right*). The top two sounds on the right are orchestral music, while the lower right is opera

Amazon EX1014 Page 747

processing, analysis/synthesis algorithms, and human perception and cognition.

One other cognition-related research area in computer music is musical style modeling, also called automatic composition. One of the leading researchers in this area is David Cope, whose experiments in musical intelligence (EMI) projects have been developed over many years now. Through musical rules and data entered by Cope in the LISP computer language, the simple analytic recombinant algorithm (SARA) program has generated imitative music in styles as varied as those of Palestrina, Bach, Brahms, Chopin, Mozart, Prokofiev, Stravinsky, Gershwin, and Scott Joplin [17.61].

Another interesting project in this area is The Brain, created by Tom Hadju and Andy Milburn of the production company "tomandandy". The Brain is partially responsible for a large amount of commercial and movie music created over the last few years, including music for *The Mothman Prophesies*, and *Mean Creek*.

17.13 Conclusions, and the Future

With the advent of cheap and powerful computers, especially laptops, the last few years have greatly democratized the tools and techniques of computer music. Many free or inexpensive sound-editing tools now include features and algorithms that were the sole domain of academic researchers only ten years ago. Further, the World Wide Web has made computer tools and music available to more people. Computer-augmented performers and performance spaces have become more common, and more commonly accepted by the public. It seems certain that the future of computer music is likely to include more developments in the areas of modeling and understanding of human perception, preference, and creativity.

References

17.1 K. Pohlmann: *Principles of Digital Audio* (McGraw Hill, New York 2000)

17.2 J.O. Smith, P. Gossett: A flexible sampling-rate conversion method. In: *Proc. Int. Conf. Acoustics Speech and Signal Processing*, Vol. 2 (IEEE, Piscataway 1984) pp. 19.4.1–19.4.2

17.3 T. Dutoit: *An Introduction to Text-To-Speech Synthesis* (Kluwer Academic, Dordrecht 1996)

17.4 R. Bristow-Johnson: *Wavetable synthesis 101: a fundamental perspective AES 101 Conference* (AES, New York 1996), Music-DSPSource Archive, (http://www.musicdsp.org)

17.5 M.V. Mathews: *The Technology of Computer Music* (MIT Press, Cambridge 1969)

17.6 R. Bracewell: *The Fourier Transform and It's Applications* (McGraw Hill, New York 1999)

17.7 X. Serra, J.O. Smith: Spectral modeling synthesis: Sound analysis/synthesis based on a deterministic plus stochastic decomposition, Comp. Music J. **14**(4), 1–2 (1990)

17.8 R.J. McAulay, T. Quatieri: Speech analysis/synthesis based on a sinusoidal representation, IEEE Trans. ASSP **34**, 744–754 (1986)

17.9 T. Verma, T. Meng: *An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio 1998 IEEE ICASSP-98* (IEEE, Piscataway 1998)

17.10 A. Moorer: The use of the phase vocoder in computer music applications, J. Audio Eng. Soc. **26**(1/2), 42–45 (1978)

17.11 M. Dolson: The phase vocoder: A tutorial, CMJ **10**(4), 14–27 (1986)

17.12 P. Cook: *Real Sound Synthesis for Interactive Applications* (AK Peters, Wellesley 2002)

17.13 P. Morse: *Vibration and Sound* (for the Acoustical Society of America, New York 1986)

17.14 K. Steiglitz: *A Digital Signal Processing Primer* (Addison Wesley, Boston 1996)

17.15 J.M. Adrien: The missing link: Modal synthesis Chap. 8. In: *Representations of Musical Signals*, ed. by De Poli, Piccialli, Roads (MIT Press, Cambridge 1991) pp. 269–297,)

17.16 H. Dudley: *The Vocoder* (Bell Laboratories Record, Mirray Hill 1939), Reprinted in IEEE Trans. Acoust. Speech Signal Process. ASSP-29(3):347-351 (1981)

17.17 B. Atal: Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Am. **A47**(65), 65 (1970)

17.18 A. Moorer: The use of linear prediction of speech in computer music applications, J. Audio Eng. Soc. **27**(3), 134–140 (1979)

17.19 D. Klatt: Software for a cascade/parallel formant synthesizer, J. Acoust. Soc. Am. **67**(3), 971–995 (1980)

Amazon EX1014 Page 748

17.20 J.L. Kelly Jr., C.C. Lochbaum: *Speech Synthesis, in Proceedings of the Fourth ICA* (Paper G42, Copenhagen 1962)

17.21 J.M. Chowning: The synthesis of complex audio spectra by means of frequency modulation, J. Audio Eng. Soc. **21**(7), 526–534 (1973)

17.22 J. Chowning: The synthesis of complex audio spectra by means of frequency modulation, Comp. Music J. **1**(2) (1977), reprinted

17.23 J. Chowning: The synthesis of complex audio spectra by means of frequency modulation. In: *Foundations of Computer Music*, ed. by C. Roads, J. Strawn (MIT Press, Cambridge 1985), reprinted

17.24 J. Chowning: Frequency modulation synthesis of the singing voice. In: *Current Directions in Computer Music Research*, ed. by M.V. Mathews, J.R. Pierce (MIT Press, Cambridge 1980)

17.25 X. Rodet: Time-domain formant-wave-function synthesis, Comp. Music J. **8**(3), 9–14 (1984)

17.26 I. Daubechies: Orthonormal bases of compactly supported wavelets, Commun. Pure Appl. Math. **41**, 909–996 (1988)

17.27 C. Roads: Asynchronous granular synthesis. In: *Representations of Musical Signals*, ed. by G. De Poli, A. Piccialli, C. Roads (MIT Press, Cambridge 1991) pp. 143–185

17.28 P. Cook: *Physically Informed Sonic Modeling (PhISM): Percussion Instruments, in Proceedings of the ICMC* (International Computer Music Association, San Francisco 1996)

17.29 P. Cook: Physically informed sonic modeling (PhISM): Synthesis of percussive sounds, Comp. Music J. **21**, 3 (1997)

17.30 C. Cadoz, A. Luciani, J.L. Florens: CORDIS-ANIMA: A modeling and simulation system for sound image synthesis. The general formalization , Comp. Music J. **17**(1), 23–29 (1993)

17.31 K. Karplus, Strong: Digital synthesis of plucked-string and drum timbres, Comp. Music J. **7**(2), 43–55 (1983)

17.32 J.O. Smith: Acoustic modeling using digital waveguides. In: *Musical Signal Processing*, ed. by Roads et.al. (Swets, Zeitlinger, Netherlands 1997)

17.33 D.A. Jaffe, Smith: Extensions of the Karplus-Strong plucked string algorithm, Comp. Music J. **7**(2), 56–69 (1983)

17.34 P.R. Cook, D. Trueman: Spherical radiation from stringed instruments: Measured, modeled, and reproduced, J. Catgut Acoust. Soc. **1**, 1 (1999)

17.35 J.O. Smith, S.A. Van Duyne: Overview of the commuted piano synthesis technique. In: *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, ed. by E. Wenzel (IEEE, New York 1995)

17.36 M.E. McIntyre, R.T. Schumacher, J. Woodhouse: On the oscillations of musical instruments, J. Acoust. Soc. Am. **74**(5), 1325–1345 (1983)

17.37 J.O. Smith: Efficient simulation of the reed-bore and bow-string mechanisms. In: *Proceedings of the International Computer Music Conference* (International Computer Music Association, San Francisco 1986) pp. 275–280

17.38 G. Essl, P. Cook: *Banded waveguides: Towards physical modeling of bar percussion instruments, in Proc. 1999 Int. Computer Music Conf., Beijing, China* (International Computer Music Association, San Francisco 1999)

17.39 MMA: *The Complete MIDI 1.0 Detailed Specification* (MIDI Manufacturers Association, La Habra 1996)

17.40 S. Jungleib: *General MIDI, A-R Editions* (A-R Editions, Middleton 1995)

17.41 M. Wright, A. Chaudhary, A. Freed, D. Wessel, X. Rodet, D. Virolle, R. Woehrmann, X. Serra: *New Applications for the Sound Description Interchange Format, Proc. Of the ICMC 1998* (International Computer Music Association, San Francisco 1998)

17.42 M. Wright, A. Freed: Open sound control: A new protocol for communicating with sound synthesizers. In: *Proceedings of the International Computer Music Conference, 1997* (International Computer Music Association, San Francisco 1997),

17.43 S.T. Pope: Machine Tongues XV: Three packages for software sound synthesis, Comp. Music J. **17**(2), 23–54 (1993), Note: these are cmix, csound, cmusic

17.44 J. McCartney: SuperCollider: A new real-time synthesis language. In: *Proc. Int. Computer Music Conference* (International Computer Music Association, San Francisco 1996) pp. 257–258

17.45 P. Burk: JSyn – A real-time synthesis API for java. In: *Proc. Int. Computer Music Conference* (International Computer Music Association, San Francisco 1998) pp. 252–255

17.46 Cook, G. Scavone: The synthesis toolkit (STK). In: *Proceedings of the International Computer Music Conference* (International Computer Music Association, San Francisco 1999) pp. 164–166

17.47 R.B. Dannenberg, E. Brandt: A flexible realtime software synthesis system. In: *Proceedings of the International Computer Music Conference* (International Computer Music Association, San Francisco 1996) pp. 270–273

17.48 R.B Dannenberg: Machine Tongues XIX: Nyquist, a language for composition and sound synthesis, Comp. Music J. **21**(3), 50–60 (1997)

17.49 G. Wang, P.R. Cook: ChucK: A concurrent. In: *On-the-fly Audio Programming Language, in Proceedings of International Computer Music Conference* (International Computer Music Association, San Francisco 2003) pp. 219–226

17.50 M. Puckette: Combining event and signal processing in the MAX graphical programming environment, Comp. Music J. **15**(3), 68–77 (1991)

Amazon EX1014 Page 749

17.51 M. Puckett: Pure data. In: *Proceedings of International Computer Music Conference* (International Computer Music Association, San Francisco 1996) pp. 269–272
17.52 D. Trueman, P. Cook: BoSSA: The deconstructed violin reconstructed, J. New Music Res. **29**(2), 120–130 (2000)
17.53 D. Trueman, C. Bahn, P. Cook: Alternative voices for electronic sound. In: *Spherical speakers and sensor-speaker arrays (SenSAs), International Computer Music Conference, Berlin, Aug. 2000* (International Computer Music Association, San Francisco 2000)
17.54 M.V. Mathews: The radio baton and conductor program, or: Pitch, the most important and least expressive part of music, Comp. Music J. **15**(4), 37–46 (1991)
17.55 T. Marin Nakra: *Inside the Conductor's Jacket: Analysis, Interpretation and Musical Synthesis of Expressive Gesture*, Ph.D. Thesis (MIT Media Lab, Cambridge 2000)
17.56 J.A. Paradiso: The brain opera technology: New instruments and gestural sensors for musical interaction and performance, J. New Music Res. **28**, 2 (1999)
17.57 J. Freeman, S. Ramakrishnan, K. Varnik, M. Neuhaus, P. Burk, D. Birchfeld: *Adaptive High-level Classification of Vocal Gestures Within a Networked Sound Instrument, Proc. Of the International Computer Music Conference* (International Computer Music Association, San Francisco 2004)
17.58 C. Chafe, S. Wilson, R. Leistikow, D. Chisholm, G. Scavone: A simplified approach to high quality music and sound over IP. In: *Proceedings of the Conference on Digital Audio Effects (DAFX)*, ed. by D. Rocchesso (University of Verona, Italy 2000)
17.59 A. Kapur, G. Wang, P. Davidson, P.R. Cook, D. Trueman, T.H. Park, M. Bhargava: *The Gigapop Ritual: A Live Networked Performance Piece for Two Electronic Dholaks, Digital Spoon, DigitalDoo, 6 String Electric Violin, Rbow, Sitar, Tabla, and Bass Guitar, New Interfaces for Musical Expression (NIME)* (McGill University, Montreal 2003)
17.60 G. Tzanetakis, P. Cook: Multi-feature audio segmentation for browsing and annotation. In: *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (IEEE, Piscataway 1999)
17.61 D. Cope: *Experiments in Musical Intelligence*, Vol. 12 (A-R Editions, Madison 1996)

Amazon EX1014 Page 750

18. Audio and Electroacoustics

This chapter surveys devices and systems associated with audio and electroacoustics: the acquisition, transmission, storage, and reproduction of audio. The chapter provides an historical overview of the field since before the days of Edison and Bell to the present day, and analyzes performance of audio transducers, components and systems from basic psychoacoustic principles, to arrive at an assessment of the perceptual performance of such elements and an indication of possible directions for future progress.

The first, introductory section is an overall historical review of audio reproduction and spatial audio to establish the context of events. The next section surveys relevant psychoacoustic principles, including performance related to frequency response, amplitude, timing, and spatial acuity. Section 3 examines common audio specifications, with reference to the psychoacoustic limitations discussed in Sect. 2. The specifications include frequency and phase response, distortion, noise, dynamic range and speed accuracy. Section 4 examines some of the common audio components in light of the psychoacoustics and specifications established in the preceding sections. The components in question include microphones, loudspeakers, record players, amplifiers, magnetic recorders, radio, and optical media. Section 5 is concerned with digital audio, including the basics of sampling, digital signal processing, and audio coding. Section 6 is devoted to an examination of complete audio systems and their ability to reproduce an arbitrary acoustic environment. The specific systems include monaural, stereo, binaural, Ambisonics, and 5.1-channel surround sound. The final section provides an overall appraisal of the current state of audio and electroacoustics, and speculates on possible future directions for research and development.

18.1 **Historical Review** 744
18.1.1 Spatial Audio History 746

18.2 **The Psychoacoustics of Audio and Electroacoustics** 747
18.2.1 Frequency Response 747
18.2.2 Amplitude (Loudness) 748
18.2.3 Timing 749
18.2.4 Spatial Acuity 750

18.3 **Audio Specifications** 751
18.3.1 Bandwidth 752
18.3.2 Amplitude Response Variation 753
18.3.3 Phase Response 753
18.3.4 Harmonic Distortion 754
18.3.5 Intermodulation Distortion 755
18.3.6 Speed Accuracy 755
18.3.7 Noise 756
18.3.8 Dynamic Range 756

18.4 **Audio Components** 757
18.4.1 Microphones 757
18.4.2 Records and Phonograph Cartridges 761
18.4.3 Loudspeakers 763
18.4.4 Amplifiers 766
18.4.5 Magnetic and Optical Media 767
18.4.6 Radio 768

18.5 **Digital Audio** 768
18.5.1 Digital Signal Processing 770
18.5.2 Audio Coding 771

18.6 **Complete Audio Systems** 775
18.6.1 Monaural 776
18.6.2 Stereo 776
18.6.3 Binaural 777
18.6.4 Ambisonics 777
18.6.5 5.1-Channel Surround 777

18.7 **Appraisal and Speculation** 778

References 778

Considering that human records extend back to cave drawings thousands of years old [18.1], it is notable that the technology to record, transmit, and reproduce sound has only existed for the last century or so. One can only

wonder at the effect it might have had on the course of history, not to mention the effect it may yet have [18.2].

Sound is hardly the easiest quantity to work with. As a vibration in air, it is invisible, dynamic, and three-dimensional (3-D), not counting the dimensions of time and frequency. Human interaction with a sound field is often dynamic: natural movements of the head and body alter the perceived sounds in a way that is difficult to predict or reproduce. Despite over a century of progress, the transducers available for recording and reproducing sound, microphones and loudspeakers, are still limited to being point-in-space devices, and are not necessarily well suited to interacting with a 3-D physical phenomenon. And yet, human beings make do with just two ears; so how hard can it be to capture and reproduce all the characteristics and nuances of what the ear perceives in nature?

The intent of this chapter is to review and evaluate the performance of audio systems based on the known performance of the human auditory system. Starting with a brief historical overview, the salient aspects of human auditory performance will be described. These in turn will be used to derive a set of general audio specifications that may be expected to apply to a broad range of audio systems. Then, some of the more common audio components will be examined in light of these specifications. Finally, the performance of complete systems will be considered, in the hope of gaining some overall perspective of progress to date and possibly fertile directions for further investigation.

18.1 Historical Review

The first golden age of audio research arguably started in the late 1800s with the inventions of the telephone by A. G. Bell and the phonograph by T. A. Edison [18.3], and ended in 1982 with the introduction of digital audio as a practical reality as embodied in the audio compact disc (CD). The inventions that inaugurated this age established that an audio channel could indeed be transmitted and recorded, as a continuous mechanical, magnetic, optical, or electrical analog of the sound vibrations, but brought with them a raft of problems, including inadequate recording sensitivity and playback volume, noise, distortion, speed irregularities such as wow and flutter, nonlinear and limited frequency response, and media wear. Although steady incremental progress was made on these problems in the decades that followed, they were all largely resolved once and for all with digital audio. Efforts since then have focused on reduction of the required data rate (audio coding) and on extending the realm of audio from single isolated channels to full three-dimensional systems. Ongoing parallel efforts to understand the physiology and psychoacoustics of the human ear more fully have continued to extend the knowledge base, and in some cases have yet to be exploited by commercial audio systems.

Why wasn't sound recording invented sooner? It is hard to say. Bell's telephone may have required a steady electric current, which presumably was not available in ancient times, but Edison's phonograph was entirely mechanical. The key to the latter invention seems to have been the recognition that sound is a vibration and that the vibration is mechanically transferable, but this notion that is readily demonstrable by putting your fingers to your lips or throat, and humming. The inventions seem to have been part of a general flourishing of physical sciences and mathematics, one that has continued to the present day. Prior to that, however, there is certainly evidence of some understanding of the acoustic processes involved, as seen in the creation of musical instruments and the design of concert halls.

Actually, both Edison and Bell were driven in part by a desire to improve telegraph systems. The telegraph was popularized by Samuel Morse, starting around 1837, although the notion of an electrical telegraph harks back at least to a demonstration of sending a simple signal over a long wire by William Watson in 1747 [18.4]. The cause was greatly helped by the invention of the electric battery by Alessandro Volta in 1799, and the storage battery in 1859 by Gaston Plante [18.5].

Many of the theoretical and mathematical foundations of early audio technologies were established in the 1800s, with work in the electrical sciences by Ohm, Faraday, Henry and others, together with research in acoustics and hearing physiology by Helmholtz, Lissajous and others. Although Edison is credited with the invention of the first device to reproduce sound, the first audio recording device is generally credited to Leon Scott, who on March 25, 1857, patented the phonautograph [18.6]. This device combined a diaphragm and a stylus to inscribe a trace of a sound on a visual medium, such as a smoke-blackened glass or cylinder, but it had no way to reproduce the recording audibly. John Tyndall seems to have followed Scott's work on visualizing sound vibrations, using a variety of different devices.

Amazon EX1014 Page 752

Alexander Graham Bell filed his patent for the telephone on February 14, 1876, just two hours before Elisha Gray filed for a similar device. Aside from its intrinsic value, the invention was significant for employing acoustical to electrical and electrical to acoustical transduction, principles that are still almost universally employed in telephony, broadcast, and audio recording.

Transduction between acoustical and electrical energy was initially less important to Edison's invention of the phonograph on December 8, 1877, since a physical medium to store the recording was necessary in any case, and there neither was nor is a ready way to directly store a dynamic electrical signal directly on such a physical medium. So it is perhaps not too surprising that the original invention employed purely mechanical transduction for both recording and playback.

With these two inventions, the foundations for analog audio were established, and the race was on to refine and commercialize the systems. In 1887, Emile Berliner patented a recording format that used lateral groove modulation on a flat disk [18.8], in contrast to Edison's use of vertical modulation on a cylinder. The underlying principles may have been much the same, but Berliner's arrangement qualified as a separate patent, and the practical aspects of mass-producing, playing, and storing disks instead of cylinders led to the grooved disk becoming the commercial recording medium of choice for over half a century.

Lacking the means to record an electrical signal directly, but seeking to avoid the wear accompanying the playback of grooved recordings with a stylus, other media were soon being explored, such as the use of magnetic recording by Valdemar Poulson in 1898 and optical recording, by Leon Gaumont in 1901. Meanwhile, the eventual foundation of the broadcast industry was begun in 1894, with the invention of the wireless telegraph by Guglielmo Marconi.

These systems were initially handicapped by a lack of viable amplification means, and although there were some imaginative attempts at mechanical amplification, the key breakthrough occurred in 1906, with Lee de Forest's invention of the *audion* vacuum-tube triode amplifier. Vacuum tubes could and would be used to amplify all manner of electrical signals, but de Forest's choice of the name *audion* seems derived directly from the word audio. It is therefore a matter of some speculation as to whether the tube amplifier would have come about when it did had it not been for the formative invention of sound recording just 29 years earlier.

Aside from resolving the problems of providing better recording sensitivity and adequate playback volume,

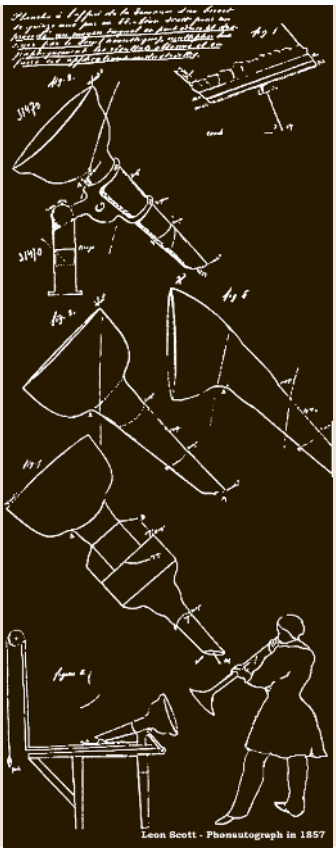


Fig. 18.1 Leon Scott and the phonautograph (After [18.7])

the vacuum tube also enabled the wireless transmission of audio, giving rise to the radio receiver and broadcast industries.

So by 1910, the basic principles, physics, and devices of audio and electroacoustics were in place, resulting in widespread continuing efforts to develop, patent, and refine components and systems, establish standards where required, and commercialize a growing variety of formats and services. Amidst competitive struggles and numerous patent fights, the following years saw a dramatic series of audio innovations, including [18.9, 10]:

- 1913: An early *talking* movie by Edison, using a cylinder synchronized with the picture.
- 1916: A superheterodyne radio circuit patented by Edwin F. Armstrong.
- 1917: A condenser microphone developed by E. C. Wente at Bell Laboratories.
- 1921: The first commercial amplitude-modulated (AM) broadcast, KDKA, Pittsburgh, PA.

Amazon EX1014 Page 753

- 1925: An electrical record cutter developed at Bell Laboratories, followed by the commercial release of 78 revolutions per minute (RPM) records.
- 1926: Iron-oxide-coated paper tape for magnetic recoding developed by O'Neill.
- 1927: *The Jazz Singer*, the first commercial sync-sound film, using audio on disk synchronized to the picture.
- 1929: The sampling theorem, the basis of digital signal processing, developed by Harry Nyquist.
- 1935: Plastic-based magnetic recording tape, from BASF.
- 1936: The first tape recording of a symphony concert, by Sir Thomas Beecham.
- 1936: A cardioid condenser microphone, by Von Braunmuhl and Weber.
- 1941: Commercial frequency-modulated (FM) broadcasting inaugurated in the US.
- 1947: The Williamson high-fidelity power amplifier developed.
- 1948: The 33 1/3-RPM long-play record, developed by CBS laboratories.
- 1951: The transistor developed at Bell Laboratories.
- 1954: The transistor radio, Sony.
- 1963: Tape cassette format developed by Philips.
- 1965: Dolby *A* noise-reduction system developed.
- 1969: Dr. Thomas Stockham experiments with digital recording.
- 1976: Stockham produces the first 16-bit digital recording, of the Santa Fe Opera.
- 1982: The digital compact disc is released commercially.

Much of the thrust of these and countless other developments in this period was to improve the quality of a given audio channel, by extending the frequency response, reducing noise, etc. As noted earlier, once digital recording came of age, these concerns were rendered somewhat moot, as adequate quality became a matter of using a sufficient number of bits per sample and adequately sampling rate to avoid audible degradation. Of course, certain parts of the audio chain, notably transducers such as microphones and loudspeakers, and associated interface electronics, have remained analog, as is likely to continue to be true at least until a direct link is established to the human nervous system, bypassing the ears altogether.

18.1.1 Spatial Audio History

Perhaps the one area of audio that was not immediately improved with the adoption of digital recording techniques was spatial reproduction. It is unlikely that the importance of the spatial aspect of audio was lost on early audio developers, nor the fact that most humans do in fact, possess two ears. It was probably just a case of priorities. Spatial audio systems would almost certainly require more channels and/or other information, which was hard to accommodate with early audio media, and the emphasis was, quite reasonably, on refining the quality of the single-channel monophonic format.

Still, there were some promising early experiments, notably a demonstration in 1881 by Clement Ader, who set up a series of microphones in front of the stage of the Paris Opera, and ran their outputs to earphones set up in a nearby room [18.9, 10]. Whether by accident or design, some of the listeners elected to use two headphones, one on each ear, and were rewarded with a crude but effective binaural spatial presentation.

Despite isolated experiments such as Ader's, there was little organized investigation of spatial audio until 1931, when Alan Blumlein single-handedly developed a range of devices and protocols to support stereo sound, including techniques to upgrade the entire audio broadcast and recording infrastructure of the time. Blumlein filed a far-ranging patent covering the bulk of his work; unfortunately, it seems to have been about a quarter century ahead of his time, as it would be the late 1950s before stereophonic sound became a commercial reality.

In the 1930s, Bell Laboratories also became interested in spatial audio, albeit with more of an emphasis on audience presentation. They concluded that stereo was ineffective for such uses, as the virtual imaging of just two channels would not work for listeners removed from the centerline. They further concluded that adding a center channel produced a far more satisfactory result, as demonstrated by Fletcher, Steinberg, and Snow in 1933.

With consumer media still limited to single-channel operation, it fell for a time to film studios to engage in occasional commercial forays into spatial audio. A notable experiment in the venue was the film *Fantasia* (1940), the road show productions of which used a synchronized three-channel soundtrack, plus control track, to feed as many as 10 loudspeakers [18.9, 10].

The use of multiple film channels accelerated somewhat in the 1950s, as film studios sought ways to differentiate their product from television. One popular approach was to use wider-than-standard 70 mm film stock, instead of the normal 35 mm, which provided enough room along the edge for several audio channels

Amazon EX1014 Page 754

on magnetic stripes. A common arrangement was three front channels plus a surround channel, the latter usually fed to an array of loudspeakers arranged around the sides and rear of the theater.

In the 1970s, the blockbuster movie genre was kicked off with the releases of *Star Wars* and *Close Encounters of the Third Kind*, along with the synergistic Dolby matrix surround system, which encodes control signals in the audio to pan two optical film channels to four output channels (again, three in front, plus surround).

In the 1990s, multichannel digital soundtracks came into use, first in movie theaters and then in home systems, generally supporting three front channels, a pair of surround channels, and a low-frequency channel, otherwise referred to as a 5.1-channel system.

18.2 The Psychoacoustics of Audio and Electroacoustics

In order to evaluate the requirements and performance of audio systems, it is necessary to establish their relation to the capabilities and limitations of the human hearing. People perceive sound on the basis of frequency, amplitude, time and, somewhat indirectly, space. The perceptual resolution limits of these quantities are fairly well established, even if the underlying mechanisms are still being explored.

18.2.1 Frequency Response

1. Range: The commonly quoted frequency range of human hearing is 20–20 000 Hz, representing the range of frequencies that can be heard as sounds. In many cases, this is a somewhat optimistic specification, and in any case certainly varies with the individual. Only young people are likely to hear out to 20 kHz. With age generally comes a reduction in the upper limit (along with other reductions in hearing acuity, a process known as presbycusis [18.11]), with typical values of the upper limit of hearing in the range of 8–15 kHz common for middle-aged adults. There has been considerable debate about the ability to hear above 20 kHz, with some claiming some sort of perception out to frequencies of 40–50 kHz. Hearing limits above 20 kHz are known to exist in some small animals. At the low-frequency end of the spectrum, a frequency of perhaps 32 Hz is typically the minimum value at which a pitch can be discerned, below which the sensation becomes more of a series of discrete vibrations. Although it is difficult to construct loudspeakers with flat response below 20 Hz, humans can still perceive such vibrations if they are sufficiently large in amplitude.
2. Sensitivity: The ear is not uniformly sensitive to all frequencies, due in part to resonances and mechanical limitations of structures in the outer and middle ear, and in part to neural effects. The sensitivity also tends to be a function of absolute loudness (defined later). Contours of perceived equal loudness were first measured by Fletcher and Munson at Bell Laboratories in the 1930s [18.12]. The figure below shows typical equal loudness curves as a function of absolute level [dB sound pressure level (SPL)].
3. Resolution: The ear exhibits a curious duality with respect to the degree it can differentiate frequencies. On the one hand, certain frequency-dependent processes, such as masking, appear to indicate

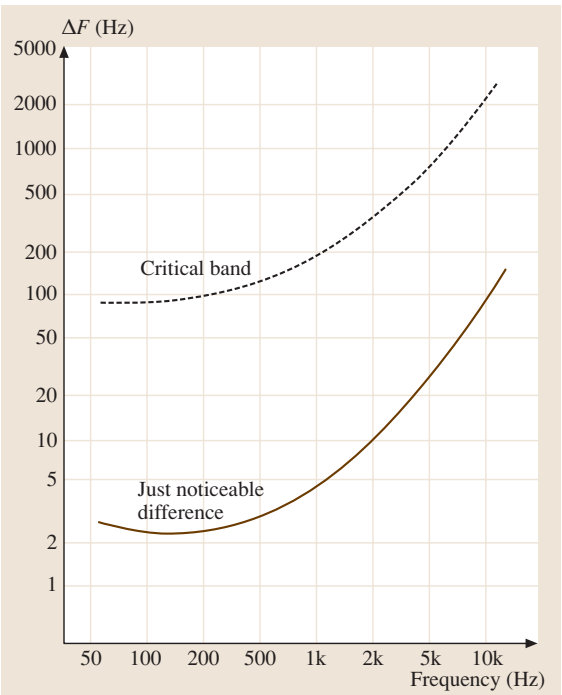


Fig. 18.2 Critical bandwidths (*upper curve*) and sine-wave pitch just noticeable difference (JND) as a function of frequency

Amazon EX1014 Page 755

a grouping of the frequency range into somewhat coarse sub-bands about 1/4 to 1/3 octave wide, called *critical bands*. On the other hand, frequency discrimination of individual sine waves on the basis of their perceived pitch has a typical resolution limit on the order of 1/100th of an octave, corresponding to a frequency change of about 0.7%. This is one of the cases where the underlying mechanisms are uncertain, and a mixture of temporal and spectral discrimination appears to be necessary to explain the various observed phenomena.

4. Pitch: In general terms, pitch is the perceptual response to differences in the frequency content of a source, but it is far more complex than a simple mapping of frequency to a perceptual scale, and can be affected by parameters such as the loudness of the signal, which ear it is played to, and the timing of the signal. The perception of pitch of spectrally complex signals is especially difficult to model, in part because a given signal may be perceived as having more than one pitch, and the signal itself may or may not actually contain spectral components corresponding to the perceived pitch [18.13].

18.2.2 Amplitude (Loudness)

Absolute Sensitivity

The total dynamic range of the human ear, that is, the range from loudest to softest sound one can perceive, is generally given to be about 120 dB a pressure range of about a million to one, or a power range of about a trillion to one. The term *sound pressure level* (SPL) is used to indicate where in this range a given sound falls, with 0 dB SPL taken as the approximate threshold of hearing at 1 kHz, corresponding to a level of about 20 μPa [18.14]. The mathematical relation between a sound pressure p and corresponding SPL value is

$$L_\mathrm{p} = \mathrm{SPL} = 20\log(p/p_0)\,,$$

where p_0 corresponds to the 0 dB SPL quoted above.

The upper limit of loudness perception is not well defined, but corresponds in principle to the onset of nonaudible phenomena such as tingling and outright pain. Depending on the listener, the upper limit is likely to be in the range of 115–140 dB SPL. (Note that exposure to such levels, especially for extended periods, can lead to hearing loss [18.11]).

As noted, sensitivity is a function of frequency and absolute level, with the greatest sensitivity observed in the range 3500–4000 Hz, mainly due to resonance of the ear canals, where the hearing threshold may actually be slightly lower in level than 0 dB SPL. The greatest variation as a function of frequency is at the lowest levels, where the threshold of hearing may rise to in excess of 20 dB SPL at high frequencies, and fully 60 dB SPL at low frequencies.

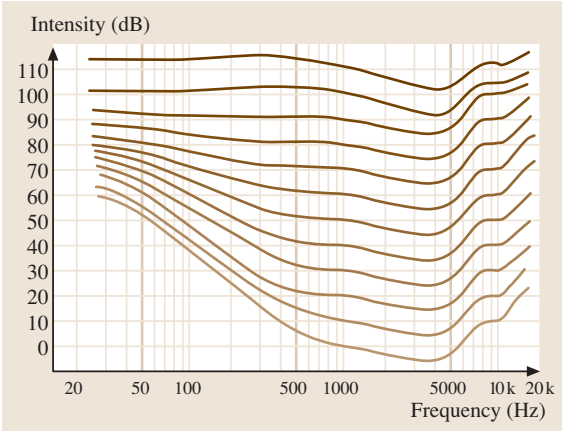


Fig. 18.3 Equal-loudness curves as a function of frequency and absolute level (After [18.9])

The variation of perceived loudness with actual sound pressure, while monotonic, is certainly nonlinear, as can be seen from the differences in spacing between the equal loudness curves at different frequencies. The actual perception of a difference in loudness will generally be less than the actual difference in SPL. According to the Sone scale of loudness perception, a difference in SPL of 10 dB at 1 kHz is roughly equivalent to a doubling of perceived loudness Sones $= 2^{[(\mathrm{Phons}-40)/10]}$ [18.15].

Because of the variation in perceived loudness with the actual SPL and frequency of a signal, it is a nontrivial matter to calculate the perceived loudness of an arbitrary, spectrally complex signal. Recent efforts using multiband analysis and power summation have improved the predictability to within close to a loudness just noticeable difference (JND) for most signals [18.16].

Resolution

The amplitude resolution of the ear is generally taken to be about 0.25 dB under best-case conditions, although for some situations it is considered to be slightly larger, on the order of 0.5–1.0 dB.

Masking

The audibility of a sine wave in isolation may be modified in the presence of other spectral components. In particular, the addition of spectral components is likely

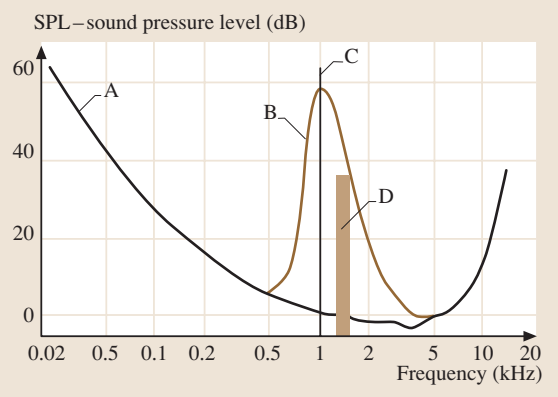


Fig. 18.4 Frequency Domain Masking: Curve A is hearing threshold with no signal present. Curve B is revised threshold in presence of 1 kHz tone C. Signal D, rising above curve A but below curve B, will be audible by itself but masked by signal C

to render an existing component less audible, a process referred to as masking.

The basic notion of masking is that a louder sound will tend to render as less audible or inaudible a softer sound that would otherwise be audible in isolation. The degree of masking will tend to depend in part on the frequency spacing of the spectral components in question. This gives rise to the notion of a prototype masking curve, delineating the threshold of audibility in the presence of a sine wave of some frequency and amplitude, as a function of frequency, as shown in Fig. 18.4.

To at least first-order approximation, the masking effect of two or more sine waves can be calculated by summing their respective individual masking curves. Since, from Fourier theory, any signal can be decomposed into a sum of sinusoids, a composite masking curve from an arbitrary signal can be calculated as the sum of the masking curves of each spectral component. This constitutes a convolution (filtering operation) of a prototype masking curve with the signal spectrum. In practice, the prototype masking curve used for the convolution is likely to vary as a function of frequency, amplitude, or signal type (noise versus sine wave). Further, nonlinearities in the ear may result in significant deviations from the predicted curve at times.

Although the masking effect of a signal is, quite logically, maximized while the signal is present, some residual of the masking effect will extend beyond the termination of the signal, and even slightly before its onset. This process is referred to as temporal masking, and typical behavior is illustrated in Fig. 18.5) [18.18].

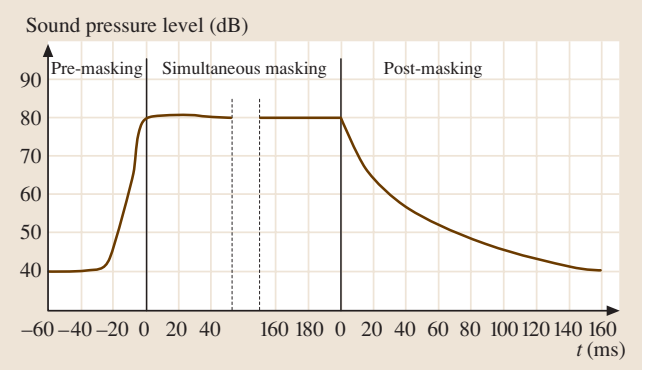


Fig. 18.5 Temporal masking characteristic

18.2.3 Timing

Because the transduction structure of the ear – the basilar membrane and associated hair cells – functionally resembles a filter bank, questions of timing must be considered as applying to the filtered output signals of that

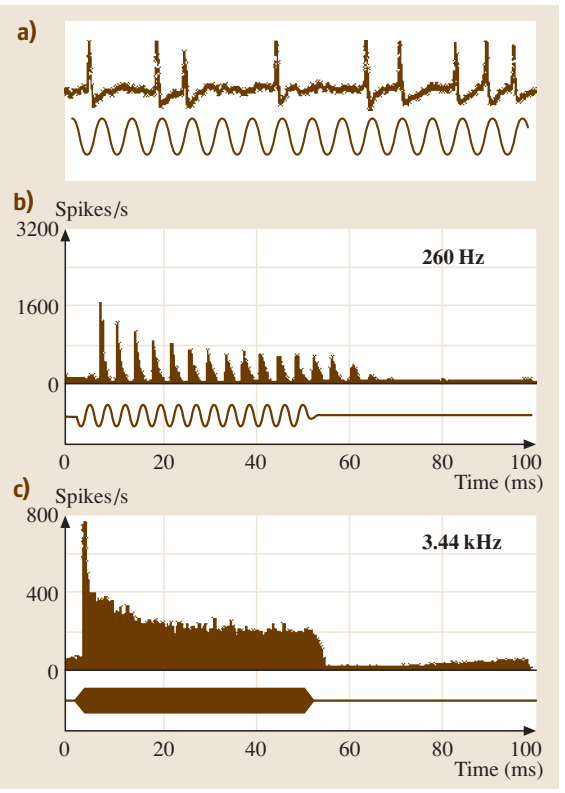


Fig. 18.6 Neural phase locking at low frequencies and lack of phase locking at higher frequencies [18.17]

filter bank. The brain does not *see* the wide-band audio signal arriving at the ear, as it might appear on an oscilloscope, but instead processes the outputs of the basilar membrane filter bank, subject to effective half-wave rectification and loss of high-frequency synchrony by the neurons (Fig. 18.6). Output neural signals from each ear are ultimately processed by multiple areas of the brain. So, it is not surprising that issues of timing are a little ill-defined, and that there are a number of time constants associated with different aspects of hearing.

For starters, there is the basic ability of the neurons in the auditory nerve to follow the individual cycles of audio impinging on the basilar membrane; i.e. to exhibit phase locking. Current research seems to put the upper frequency of this activity at about 5 kHz, although this seems somewhat at odds with lateralization of sine waves on the basic of interaural time difference, which only extends up to about 1500 Hz.

The ability to follow individual cycles of audio may aid in pitch perception, as the effective shapes of the basilar membrane filters appear too broad to account for the observed pitch resolution. However, the physiological mechanisms for how this might be accomplished are still the subject of some debate.

Other time constants appear to apply to the aggregate audio, regardless of the presence of a filter bank at the input. Already noted are the masking time constants, which substantially extend only 1–2 ms prior to the onset of a loud sound, but continue for several tens of milliseconds after its cessation.

Each of these may be related to a more fundamental time constant. The short time constant of about 1–2 ms may represent the shortest time one can perceive without relying on spectral cues, while the post-masking interval may be related to the fusion time. The fusion time in particular seems to represent a kind of acoustic integration time, or the limit of primary acoustic memory, typically on the order of 30–50 ms, and appears to be associated with the lower frequency limit of the audio band (20–30 Hz). It also seems to explain why echoes are only heard as such in large cathedral-like rooms, as they are integrated with the direct arrivals in normal-size rooms.

Table 18.1 Summary of approximate perceptual time/amplitude limits

Parameter	Range	JND
Amplitude	120 dB	0.25 dB
Premask	N/A	2 ms
Postmask	N/A	30–50 ms
Binaural timing	700 µs	10 µs

On the other hand, binaural timing clearly exhibits higher resolution, as differences of interaural timing on the order of 10 ms may be audible.

18.2.4 Spatial Acuity

Strictly speaking, the spatiality of a sound field is not perceived directly, but is instead derived from analysis of the physical attributes already described. However, given its importance to audio and electroacoustics, it is nonetheless useful to review the processes involved.

Specification of the position of a sound source relative to a listener in three-space requires three coordinates. From the mechanisms used by the human ear, the natural selection is to use a combination of azimuth, elevation, and distance. Of these, distance is generally considered to be an inferred metric, based on the amplitude, spectral balance, and reverberation content of a source relative to other sources, and is not especially accurate in any absolute sense.

This leaves directional estimation as the primary localization task. It appears that the ear uses separate processes to determine direction in a left/right sense and a front/top/back sense. Given the physical placement of the ears, it is not surprising that the left/right decision is the more direct and accurate, depending on the interaural amplitude difference (IAD) and interaural time difference (ITD) localization cues.

Although the ear is sensitive to IAD at all audible frequencies, as can be verified by headphone listening, the head provides little shadowing effect at low frequencies, since the wavelengths become much longer than the size of the head, so IAD is mostly useful at middle to high frequencies.

The JND for IAD is about 1 dB, corresponding to about 1 degree of an arc for a front/centered high-frequency source. As a sound moves around to the side of the head, the absolute IAD will tend to increase, and the differential IAD per degree will decrease, causing the angular JND to decrease accordingly.

The ITD lateralization cue has a rather astonishing JND of just 10 ms, for a source on the centerline between the ears, corresponding to an angle of about 1 degree. For signals off to one side, the ITD will typically reach a maximum value of just 750 ms – 3/4 of a millisecond. As with the IAD cue, the sensitivity of the ITD cue declines as a source moves around to the side of the head.

Also like IAD, this cue is usable at all frequencies, albeit with a key qualification. At frequencies below about 1500 Hz, the neural response from the basilar

Amazon EX1014 Page 758

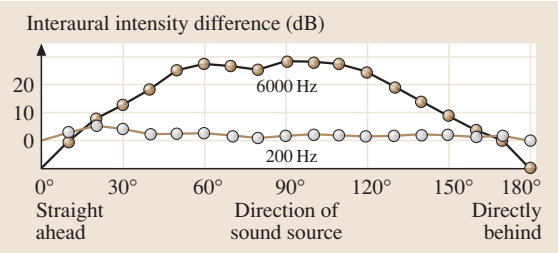


Fig. 18.7 Interaural amplitude difference as a function of direction [18.19]

membrane remains phase locked to the bandpass-filtered audio waveform. At such frequencies, even a modest amount of interaural phase shift of a sine wave can result in a shift in the position of the audio image. Above 1500 Hz, neural phase synchrony is effectively lost, at least as far as ITD perception is concerned, so the interaural phase of a high-frequency sine wave has little or no perceptual effect on its apparent position. However, the brain can still make fairly accurate judgments of the ITD of the *envelopes* of the high-frequency bandpass-filtered signals, assuming there is sufficient variation in the filtered envelopes. This will tend to make spectrally dense high-frequency signals easier to localize than sparse signals, since the latter may only exhibit constant-envelope sine waves after the critical band filtering of the basilar membrane.

While localization in the horizontal direction is firmly mediated by two strong, well-behaved cues, localization in the vertical dimension is decidedly less well behaved. The goal of vertical localization is to differentiate the positions of sources on locus of constant IAD/ITD. For sources equidistant from the two ears, the locus is a plane passing though the center of the head, referred to as the median plane. For sources off to one side, the locus is a cone, sometimes referred to as the *cone of confusion*. The best-case resolution, for front-centered sources, is typically on the order of 5°, which is quite a bit less precise than the horizontal resolution of about 1° under the same conditions. As with horizontal localization, vertical resolution appears to diminish for side sources, although since the cone of confusion becomes progressively narrower, the net spatial error resulting from a vertical localization error is likely to diminish. For sources at the side, the cone of confusion collapses to a line.

One demonstrated vertical localization cue [18.20] is generated by the interaction of the arriving sound and the folds of the pinnae. The pattern of reflections off the folds depends on the direction of arrival of a sound, although not in any simple, monotonic fashion. One consequence of these pinna reflections is the introduction of sharp peaks and dips in the frequency response, which the brain can apparently learn to associate with specific directions. Because of the size of the pinna, and resulting reflection delays, these direction-dependent response delays only appear at high frequencies, above about 7 kHz, and so are less effective for low-frequency or narrow-band signals. Vertical localization of signals below 7 kHz may be possible on the basis of pinna cues if the brain is performing some sort of autocorrelation of each ear's signal, and can detect the delays in the time domain, but so far there is little evidence to support this hypothesis.

A second vertical localization cue is dynamic in nature, namely the modulation of the horizontal localization cues – ITD and IAD – with movement of the head, which will vary with the vertical position of the source. Given a front source on the median plane, for example, a rightward movement of the head will result in a left ear arrival that is louder and sooner than the right ear arrival. For a rear source, the correspondence will be opposite, and for a top source, there will be essentially no change in the ITD/IAD values. By correlating head movement with these sonic changes, the brain can deduce the vertical position of the source.

Finally, the presence and pattern of room echoes, once learned, and in concert with the above cues, can reinforce the vertical position of a source, especially whether in front of or behind the listener.

Even so, the vertical localization cues tend to be somewhat individual, like fingerprints, and, for example, playing a recording made with microphones placed in the ear canals of one subject may not result in good vertical localization for another listener.

18.3 Audio Specifications

With the basic limits of auditory resolution now reviewed, consideration can be given to the requirements for high-quality audio systems in terms of those limits. While individual audio components can be expected to exhibit correspondingly individual performance requirements, as will overall systems, there are some

Amazon EX1014 Page 759

specifications that can be expected to apply to all elements of the signal chain. Some of these are considered traditional audio specs that have been used as general metrics for many years. Happily, in at least some cases, modern audio systems are of such high quality as to render some specifications moot. In other cases, complete fulfillment of some of the requirements is something that may never be completely realized.

18.3.1 Bandwidth

It is reasonably evident that, given the nominal 20–20 kHz range of human hearing, it will be desirable for audio systems with aspirations to sonic accuracy to support at least that frequency range. Since an audio signal may pass through any number of processors on its way from acquisition to ultimate reproduction, it is generally advisable for each processor to have a wide-enough pass band to avoid progressive bandwidth limitation.

For most audio devices during the formative years of audio development, this requirement was daunting, and rarely achieved outside the laboratory, if then. There were too many mechanical vibrating structures in the chain, and it is very difficult to make them respond consistently over a 1000:1 frequency range. Grooved disc recordings were limited by large-diameter, low-compliance styli, and slow inner-groove velocities of constant-RPM discs. Optical media, like film, and magnetic media, were both limited by linear speed, and the size of the effective apertures used for recording and playback. The linear speed of such media was chosen as a compromise with playing time, and the smallness of the apertures was limited by high noise levels and diminishing signal level. Broadcast media were limited by bandwidth constraints and noise.

Not counting a straight piece of wire, amplifiers are the audio component that have probably had the easiest time achieving full audio bandwidth, particularly small-signal amplifiers, thanks in part to the relatively low mass of the electron. Power amplifiers have had a slightly harder time, partly because devices that handle high powers often come at the expense of compromises in other parameters, and partly because loudspeaker systems often exhibit ill-controlled, non-resistive loads. The response and linearity of amplifiers of all sorts was materially improved by the invention of negative feedback by Harold Black of Bell Laboratories, on August 2, 1927 [18.21].

An early recording system with fairly wide bandwidth was demonstrated on April 9–10, 1940, by Harvey Fletcher of Bell Laboratories, and conductor Leopold Stokowski. This employed three optical audio tracks on high-speed film, with a fourth track for a gain control signal, apparently implementing a wide-band compressor/expander for noise reduction, and possessing a frequency response of about 30–15 000 Hz.

It was probably not until the mid-1950s that such a bandwidth became commonplace in consumer media, with improvements such as the long-play record, commercial FM broadcasting, and refinements in magnetic tape, heads, and circuits.

As with a number of other basic audio specifications, the question of full-bandwidth frequency response was, to a great extent, rendered moot with the commercial introduction of digital audio in 1982. As discussed in a subsequent section, the bandwidth of a digitally sampled signal is largely dependent on the sampling rate chosen, which must be at least twice as high as the highest audio frequency of interest. Thus, conveying a bandwidth of 20 kHz requires a sampling frequency of at least 40 kHz. Current standard sampling frequencies in common use for full-bandwidth audio systems are 44 100 Hz and 48 000 Hz, depending upon the medium. Sometimes, a slight reduction in high-frequency response is tolerated, and a sampling frequency of 32 000 Hz is used to reduce the data rate slightly. Audio perfectionists, on the other hand, often prefer sampling rates of 96 000 Hz or more, to support an effective bandwidth in excess of 20 kHz and thereby provide additional safety margin to avoid alias distortion, preserve harmonics above 20 kHz, avoid possible intermodulation distortion, and mitigate possible filter distortion.

Once converted to digital representation, an audio signal tends to maintain its bandwidth unless explicitly limited by some digital signal processing (DSP) operation.

Despite the advances of digital audio, full-range audio systems are still the exception. The chief culprits for this situation are the transducers, especially loudspeakers. At low frequencies, naturally occurring combinations of mass and compliance lead to a fundamental resonance for most typical loudspeaker drivers that falls well inside the audio band, imparting a high-pass characteristic that usually rolls off at lower frequencies at a rate of at least 12 dB per octave. At high frequencies, a similar fundamental resonance leads to a similar roll-off, with a lowpass characteristic. Some of these roll-offs can be partially mitigated with electronic equalization, but the steepness of the characteristic makes it difficult to extend the response very far

Amazon EX1014 Page 760

with equalization, without incurring unduly high-power requirements.

In general, most musical and speech energy is concentrated in a subset of the audio band, say from 100 Hz to 10 kHz, allowing the majority of audio signals to be satisfactorily reproduced with modestly compromised system bandwidth, to the benefit of size, cost, and bulk.

18.3.2 Amplitude Response Variation

From the basic amplitude acuity of the ear, it is logical to require that the pass-band deviation in frequency response be less than an amplitude JND, say within ± 0.25 dB of flat. Of course, this is the desired net response of an entire signal chain, and if deviations are known to exist in one part of the chain, it may be possible to apply compensating equalization elsewhere in the chain.

As with the quest for full audio bandwidth, early audio systems for many years had trouble meeting this requirement, in part from mechanical limitations. Amplifiers again have traditionally had the easiest time meeting this specification, and modern-day DSP is limited only by the precision of the analog/digital converters, which is generally quite good in this regard, plus any deliberate signal-processing spectral modifications.

The worst offender is many cases is the listening room, where echoes and resonances are likely to make nearly any sine-wave frequency response plot a jumble of peaks and dips covering many dB. The three-dimensionality of the room and the fact that echoes arrive from diverse directions work to reduce the audibility of many of these measured deviations.

After rooms, loudspeakers have the hardest time meeting this flat-response requirement, being mechanical devices and having to handle a lot of power and move a lot of air. In order to cover the bulk of the audio band, loudspeaker designs typically employ multiple drivers, each optimized for a portion of the audio band, but this in turn introduces potential response deviations from required crossover networks.

In considering reasonable values for most audio specifications based on psychophysical performance, it is prudent to consider both monotic and dichotic performance; that is, each ear considered in isolation, and then the binaural performance of the two ears together. In the case of amplitude response variation, similar JNDs apply, so there is no compelling reason to differentiate the situations in this case.

18.3.3 Phase Response

In the course of passing through an audio component, an audio signal may encounter delays, which may be frequency dependent, and may therefore be categorized as either frequency-dependent time delays, group delays, or phase shifts. It is therefore prudent to establish preferred perceptual limits on the allowable deviations of these quantities.

Of course, a pure wide-band time delay will be manifested as a linear change in phase shift proportional to frequency, so to separate the effects of time delay and phase shift, so the latter ("frequency-differential single-channel phase shift") may commonly be understood to represent the deviation from a linear-with-frequency phase characteristic.

It is a little difficult to be completely definitive about a phase-shift specification, in part because there are multiple neural timing mechanisms involved. For a sound channel considered in isolation, phase perception is rather weak. There does not seem to be an explicit means of comparing the fine-grain phase or timing of signals at significantly different frequencies (more than a critical bandwidth), at least for steady-state signals. This is sometimes used as the basis of a claim that the ear is virtually phase deaf, although that is probably an overstatement. Still, if one constructs a steady-state signal with widely space spectral components that are in slow precession, say a combination of 100 Hz and 501 Hz (360° per second relative phase precession), the resulting percept is almost entirely constant.

So under what conditions are phase shifts audible? There seem to be at least three such mechanisms: fusion time, transient fusion time, and critical band envelopes.

As noted earlier, fusion time refers to the audio event integration time of the ear, generally in the range of 20–50 ms. Separate audio events that occur farther apart than the fusion time will generally be heard as separate. If a spectrally complex signal is put through a system where some frequencies are phase- or time-shifted by more than the fusion time, the signal may temporally defuse, rendering the phase shift quite audible. (A 20 ms phase delay at 30 Hz corresponds to 216 degrees of phase shift, not an unrealistically large value.) This can occur, for example, if a sharp pulse is passed through a third-order, 1/3-octave filter bank tuned for flat magnitude response. What starts out as a click will emerge as a "boing", quite audibly distinct from the source.

Transient fusion time refers to the shortest interval the ear can differentiate temporally, as exemplified by the premasking interval of about 1–2 ms. If a sharp tran-

Amazon EX1014 Page 761

sient, shorter than the transient fusion time, is passed through an audio channel where the high-frequency phase shift exceeds the transient fusion time, the output signal may sound audibly time-smeared compared to the source. Although the effect can be subtle, it is often manifested as the difference, e.g., between a sharp strike with a hard mallet, and a wire-brush mallet hit.

Critical band envelopes come into play with regard to the audibility of phase shifts with steady-state spectrally dense signals, more specifically signals in which two or more significant signal components fall within a common critical band. Since the basilar membrane filter bank will be unable to segregate these signals into separate bands, they will beat with one another, and the pattern of beats may be quite audible, for example, a complex composed of equal parts 500 Hz and 501 Hz, which can be considered the 500th and 501st harmonic of a 1 Hz pulse train. These two spectral components are certainly within a critical band of one another, and the instantaneous phase relation between them will certainly be audible; indeed the signal will cancel altogether once per second. Compare that to the earlier example of 100 Hz and 501 Hz, the 100th and 501st harmonic of a 1 Hz pulse train, where the phase relation remained virtually inaudible because of the wider frequency spacing.

So, it can be safely hypothesized that keeping frequency-differential single-channel phase shift under 1 ms is probably pretty inaudible, assuming small changes over the width of a critical band, and in practice, with most non-transient music, speech or other natural sounds, it can be expected that phase shifts corresponding to frequency-differential time shifts much under the fusion time of 20–50 ms (at least at low frequencies) will be fairly benign, if not almost totally inaudible under most conditions.

Once again, amplifiers have the easiest time meeting this requirement, with multi-driver loudspeaker systems being one of the more problematic components. Although sharp resonances, the bane of mechanical acoustic components, may exhibit audible amounts of phase shift, they will often be rendered insignificant compared to the amplitude deviations introduced by such resonances. And, again, with DSP systems, audible phase shift is not likely to be an issue as long as the signal remains in the digital domain, unless such phase shift is purposely introduced with signal processing.

As can be readily gleaned from binaural psychophysics, dichotic phase shift – the phase shift between a pair of channels – can be more audible at lower delay levels than monotic phase perception. A timing difference of even a single sample at 48 000 Hz sampling, about 20 μs, is twice the JND of this parameter. Of course, the perceptual metric for binaural phase difference is a change in apparent image position or width, rather than the spectral or amplitude change associated with monotic phase shift. It is possible for audio systems to have audible amounts of monotic phase shift and still meet the dichotic (interchannel) phase requirement, as long as all channels have the same amount of phase shift.

As usual, amplifiers and digital signals have the easiest time with this specification. In the past, some multichannel analog recording media, such as magnetic tape recorders or optical film recorders, tended to exhibit small amounts of interchannel timing error, owing to misalignment of the recording and playback apertures. Real-world rooms and loudspeakers have the hardest time, since differential timing of 10 μs corresponds to a path-length difference of only about one eighth of an inch, and few listeners are willing or able to hold their heads that still. This effect is somewhat mitigated because the plethora of echoes in a typical room leads the brain to discount timing information in favor of first-arrival amplitudes, which are somewhat better behaved. Still, it can be readily appreciated that the virtual imaging of stereo reproduction is likely to be fairly sweet-spot dependent, since listener positions near a loudspeaker will tend to receive sound both sooner and louder from the near loudspeaker, resulting in spatial distortion of the sound stage.

18.3.4 Harmonic Distortion

The presence of nonlinearities in the transfer function of an audio device or system is likely to have the effect of introducing new, spurious spectral components. For a single isolated sine wave, the nonlinearity will result in a new waveform with the same periodicity as the original signal, which can then be expressed via Fourier analysis as a sum of the original sine wave plus harmonics thereof [and perhaps a direct-current (DC) component]. For more spectrally complex signals, each discrete spectral component will be subject to such spurious harmonic generation. This process is referred to as harmonic distortion, and the power sum of the generated harmonics, divided by the power of the original signal, is referred to as total harmonic distortion (THD), usually expressed as a percentage.

The limit of acceptable quantities of THD is the point at which the generated harmonics become audible. In part, this may be signal dependent, since distortion products of a low-level signal may themselves fall be-

Amazon EX1014 Page 762

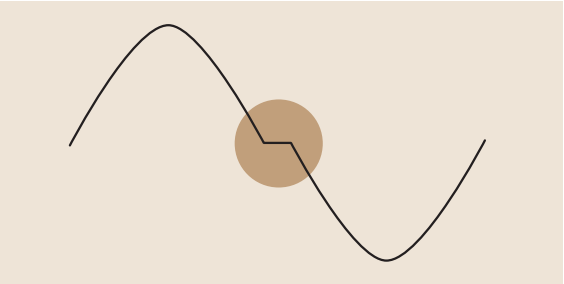


Fig. 18.8 Sine wave with crossover distortion (http://www.duncanamps.com/technical/ampclasses.html)

low the threshold of hearing, rendering them inaudible. For higher-level signals, the distortion products will be inaudible if they are masked, rendering the effect of harmonic distortion decidedly signal dependent. (The amplitude dependency may by further emphasized by nonlinearities that are themselves amplitude dependent, such as clipping.) For example, a wide-band noise signal may mask all harmonic distortion components of a moderate level. Harmonic distortion may therefore be most audible with spectrally sparse signals, with minimal masking components, such as a single sine wave.

For a sine-wave signal, the maximum acceptable amount of harmonic distortion can be deduced from the masking curve of that sine wave. From the masking curve shown in Fig. 18.4, it may be deduced that, since the masking level for a 1 kHz tone is around −35 dB at 2 kHz, the tolerable amount of second-harmonic distortion is about 2%. By 3 kHz, the masking curve is down around 50 dB, or about 0.3% third-harmonic distortion. Higher-order harmonics would benefit from even less masking, making them even more likely to be audible, until their amplitude falls below the threshold of hearing. Thus, the audibility of harmonic distortion will depend on the order of the distortion, with low-order distortion generally being less audible than high-order distortion.

Low-order harmonic distortion is generally associated with smooth, gentle nonlinearities, such as the typical onset of clipping of a vacuum-tube amplifier or a loudspeaker driver. High-order harmonic distortion, in contrast, is likely to result from sudden, sharp nonlinearities, such as the crossover distortion of an improperly designed solid-state power amplifier.

Among more traditional audio devices, playback of grooved media was probably fraught with the greatest amount of high-order THD, caused by sharp mistracking of the stylus. In the age of digital audio, low-level undithered digital signals will likely exhibit stair-step behavior which will also tend to be rich in high harmonic content, rendering them highly objectionable from the perspective of THD. Although this can lead to a desire for very high quantization accuracy (say 32-bits per sample), proper use of dither to spread the harmonic energy out as noise is generally the preferred solution.

18.3.5 Intermodulation Distortion

Intermodulation (IM) distortion results when two or more spectral components interact with a nonlinearity. The resulting distortion products will tend to be generated at sums and differences of the original spectral components. For even low-order IM distortion, these distortion products may be far in frequency from the source signal components, and may in fact be much lower in frequency. Since masking tends to be most effective when the signal to be masked is higher in frequency than the masker, having IM distortion products lower in frequency than the generating components renders them more likely to be audible, in turn making IM distortion potentially more bothersome than THD.

Serious IM distortion tends to be associated with sharp nonlinearities, and was a problem with analog grooved recordings when the stylus mistracked, as well as with poorly designed power amplifiers exhibiting crossover distortion. It can also show up with low-level signals and digital converters. Loudspeakers usually have less trouble with IM distortion, in part because any nonlinearities are usually more gentle (associated with maximum output limiting), and in part because IM distortion products located outside the range of a driver will tend not to be radiated.

Although the theoretical audible limit of IM distortion is a miniscule fraction of 1%, a value of 0.1% IM distortion or less is usually pretty safe.

18.3.6 Speed Accuracy

Considerations of speed accuracy encompass both absolute accuracy and dynamic variations, sometimes subdivided into slow variations called *wow* and faster variations called *flutter*. Except in extreme cases, timing error is usually perceived as an error in pitch, rather than time. Unless it is drastically in error, absolute speed accuracy is likely to be of concern mostly to folks with absolute pitch.

Speed variations such as wow and flutter are likely to be audible to anyone with normal pitch perception. Given a pitch JND of 0.7%, which would correspond to peak-to-peak just-noticeable speed variation, the cor-

Amazon EX1014 Page 763

responding root-mean-square (RMS) speed variation would be about 0.25%. Since speed variation may be cumulative, a preferred figure of 0.1% is likely to be specified [18.22]. There may also be cases in which cancellations from room echoes cause amplitude variations resulting from flutter that may require still lower limits of flutter to be rendered inaudible.

The analog mechanical and electromechanical recording devices of the past all have had to contend with speed accuracy. Turntables, tape drives of all stripe, and film projectors all fought off the effects of speed variation, usually with large flywheels and elaborate bearings. Getting a small lightweight tape drive like the Sony Walkman® to have low flutter was a good engineering trick. Properly designed and maintained equipment was generally able to meet the required speed accuracy specification.

In the age of digital audio, speed errors of all sorts are mostly a dead issue. Sampling clocks controlled by quartz oscillators provide speed accuracy and steadiness far in excess of what the ear can hope to detect.

18.3.7 Noise

Noise is a first cousin to distortion in its characteristics and considerations. It is a random signal, hopefully low in level, that gets tacked onto an audio signal in the course of almost any processing, analog or digital, and it is generally cumulative in nature. It often sounds like a hiss. Noise will be inaudible if it is either below the threshold of hearing, or masked by the audio signal.

Sources of Noise

Noise sources include:

Acoustical: Random fluctuations of air molecules against the eardrum or a microphone diaphragm
Electrical: Tiny fluctuations of electrical charge in analog electronic circuits from discrete electrons
Mechanical: Surface irregularities of grooved media
Magnetic: Irregularities of magnetic strength from discrete magnetic particles in recording tape
Arithmetic: Random errors from DSP quantization of numerical signals to discrete values.

There are other common spurious signals, such as hums and buzzes, that can corrupt an otherwise pristine audio signal, and although strictly speaking they may not be considered noise, they are generally at least as undesirable.

In any case, the level of noise can be considered as a function of frequency, giving rise to the notion of the spectrum or spectral density of the noise. If the value of an audio noise signal at any point in time is uncorrelated with the value of the noise at any other point in time, the signal is an example of what is called *white noise*. Spectrally, white noise is said to be flat in a linear frequency sense. That is, in essence, the output energy of a bandpass filter of bandwidth B Hz. supplied with white noise, will be the same regardless of the center frequency of the filter (providing, of course, that the pass-band of the filter is substantially within the audio band).

Because the filters on the basilar membrane are approximately logarithmic (not linear) in center frequency and bandwidth, and because they tend to integrate the signal energy within each pass-band, white noise will likely be perceived as having a rising high-frequency spectrum, even though it is flat on a linear frequency basis. In other words, as the center frequency of each effective basilar membrane bandpass filter increases, so too does its bandwidth, roughly in proportion, ergo the total filtered energy of a white-noise source will increase with frequency, leading to the bright, top-heavy perception. For each successive octave, or doubling of center frequency, the bandwidth will double, ergo so too will the bandpass white-noise signal energy, leading to a characteristic spectral rise of 3 dB per octave.

Note: dB = 10 log (ratio of power) = 20 log (ratio of amplitude).

If white noise is processed with a filter having a constant roll-off characteristic of −3 dB per octave, the output noise will have constant energy per octave, or fraction thereof, and will sound flatter to the ear. Such a signal is referred to as pink noise.

Although many forms of noise are independent of the presence of an actual signal, there are mechanisms whereby noise, or noise-like artifacts, are signal dependent. The most common such behavior is for the level of the noise to be roughly proportional to signal level. This type of artifact can be especially pernicious, as the dynamic behavior tends to call attention to the noise.

Although there are numerous ways of dynamically suppressing noise if it is unavoidable, it is by far preferable to avoid it in the first place as much as possible via proper low-noise design, proper shielding and grounding, etc.

Amazon EX1014 Page 764

18.3.8 Dynamic Range

Dynamic range is the amplitude range from the largest to the smallest signal an audio device or channel can handle. In cases where there is no minimum signal, other than zero level, the minimum useable level may be taken as the noise level of the channel.

If the noise level of the channel is substantially independent of signal level, the dynamic range of the channel may alternately be referred to as the signal-to-noise ratio (SNR). If the noise is signal dependent, the audibility of the noise will depend on the instantaneous signal-to-noise ratio, which is likely to be quite different from the dynamic range.

Since both the noise level and the overload point of a channel are likely to be functions of frequency, so too will the dynamic range. A common characteristic of the noise of traditional analog recording systems has been rising noise levels at the ends of the audio band, coupled with diminishing maximum signal levels, a double whammy that tends to limit the resulting dynamic range, in some cases rather severely.

Historically, unadorned analog channels have tended to have overall dynamic-range specifications on the order of 45–60 dB. This is true of such media as FM stereo, grooved records, optical film soundtracks, and magnetic tape. Perhaps the most limited hi-fi medium in this regard has been cassette tape, which manages about 55 dB at midband and on the order of 30 dB dynamic range at high frequencies.

Happily, DSP systems usually have flat overload characteristics and flat, or at least white, noise floors. The dynamic range of a DSP system is largely a question of the number of bits used to represent the signal. A rough rule of thumb is that each additional bit in the representation of sample values adds 6 dB to the dynamic range. The common 16-bit format used on digital compact discs yields a theoretical dynamic range of 96 dB. In practice, the combined effect of real-world imperfections and converter noise limits the practical attainable dynamic range with 16-bit audio to a little short of 96 dB, perhaps about 93 dB. A 24-bit converter can approach 144 dB dynamic range, which not only allows coverage of the full human auditory dynamic range, but allows margin for error in setting recording levels and subsequent mixing and processing operations.

18.4 Audio Components

With the common audio specifications now delineated, consideration can be given to some of the common audio components, their typical specifications, and specific characteristics and limitations.

18.4.1 Microphones

For any acoustical recording, the type, pickup pattern, position, and characteristics of the microphone are likely to have the greatest effect on the resulting quality of the recording. A microphone is, of course, an analog transducer, generally used to convert acoustic signals to proportional electrical signals. It is the functional analog of the eardrum in the human auditory system.

The term microphone was coined by Sir Charles Wheatstone in 1827, although at that point it referred to a purely acoustical device, such as a stethoscope. Its start as a transducer was pretty much coincident with Bell's invention of the telephone. Bell's initial *transmitter* used a vibrating needle connected to a diaphragm and immersed in an acid bath to implement a variable resistance responsive to acoustic waves [18.23]. This was followed in short order by more practical, rugged designs by Bell, Edison, Berliner, Hughes, and others.

Common to virtually all microphones ever made is a diaphragm. Most microphones do not directly sense the instantaneous pressure deviations of an airborne acoustic wave. Instead, the sound waves impinging on a diaphragm cause movement of the diaphragm, and it is this movement that is sensed and converted to an electrical audio signal.

The methods used for the mechanical-to-electrical transduction are many and varied. Bell's telephone patent seems to describe an odd type of electromagnetic microphone, in which the diaphragm was mechanically coupled to an armature in a magnetic circuit. Acoustic vibrations caused vibration of the armature, in turn inducing changes in the reluctance of the magnetic circuit, which induced a dynamic current in a coil of wire.

Later dynamic microphones simply attached a fine coil of wire to the diaphragm, and suspended it in a magnetic field. Movement of the diaphragm induces a voltage in the coil of wire, presumably proportional to the velocity of the diaphragm, and said voltage can

then be amplified as necessary. Dynamic microphones do not require an external source of power. Their popularity was limited prior to World War II by a lack of powerful permanent magnets.

Many early telephone transmitters used carbon in one form or another, typically carbon granules coupled to the diaphragm such that vibration of the diaphragm would dynamically squeeze the granules, causing a variation in resistance that could be electrically sensed by an applied voltage. This arrangement could be scaled up in size and voltage enough to produce a sizeable signal, of particular importance prior to the advent of electrical amplification.

Another type of self-generating microphone is the crystal microphone, based on the piezoelectric effect discovered by the Curies in 1880. Here, the diaphragm constricts a thin crystal made of Rochelle salt, ceramic, or similar material, directly producing a voltage. The low compliance of the piezoelectric element typically limits the performance of such a microphone, especially the extent of its frequency response.

Indeed, the performance of most of the microphone types mentioned so far is somewhat limited by the need to attach some object or structure to the diaphragm to detect its motion, increasing the mass of the moving elements, which in turn tends to limit the high-frequency response.

Two types of microphones avoid any mechanical attachment to the diaphragm: the ribbon microphone and the condenser microphone.

A popular variant of the dynamic microphone, the ribbon microphone, uses a thin metal ribbon diaphragm suspended in a magnetic field. Movement of the diaphragm induces a current in the diaphragm, which can then be amplified. In effect, the diaphragm acts as a single-turn moving coil. The ribbon microphone was developed by Schottky and Gerlach in 1923, and considerably refined by Harry Olsen of RCA in 1930 [18.26].

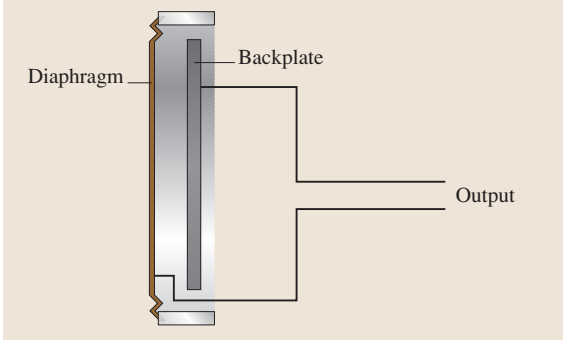


Fig. 18.9 A condenser microphone consists of a very thin diaphragm suspended parallel to a backplate (After [18.24])

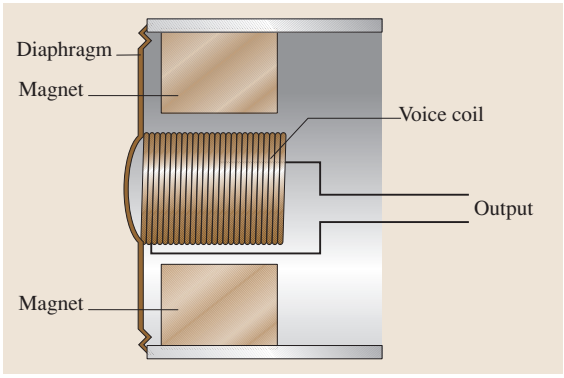


Fig. 18.10 A dynamic microphone uses a coil of wire attached to a diaphragm, suspended in a magnetic field (After [18.24])

The condenser microphone is based on the notion of electrical capacitance. A metal, or other conducting, diaphragm is placed close to a second, fixed conducting plate, forming a capacitor, and a polarizing voltage is applied between the plates. Movement of the diaphragm causes alterations in the capacitance, in turn causing corresponding alteration of the charge on the plates, which is converted to a dynamic voltage by a resistor, and amplified as necessary. The invention of the condenser

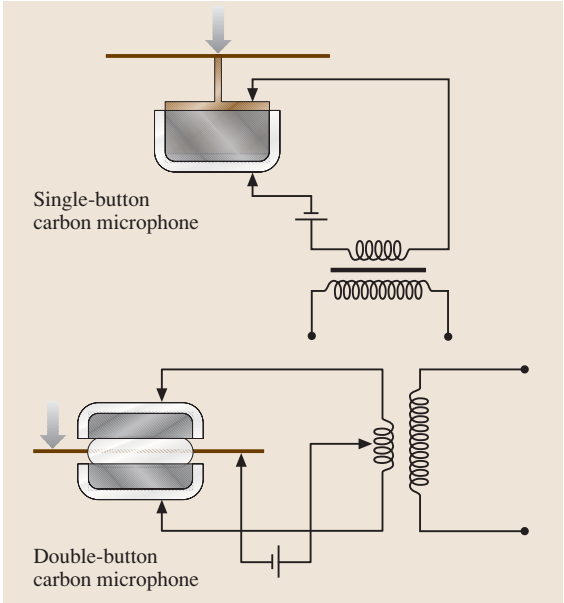


Fig. 18.11 A carbon microphone (After [18.25])

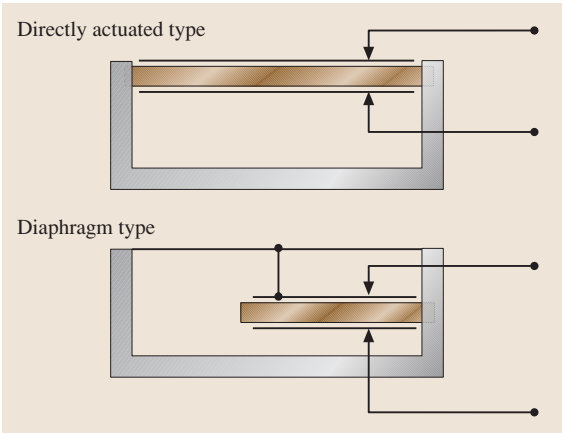


Fig. 18.12 Piezoelectric microphones (After [18.29])

microphone is generally credited to E. C. Wente of Bell Laboratories, in 1917 [18.27]. A popular modern variant is the electret condenser microphone, perfected in 1962 by Sessler and West of Bell Laboratories [18.28]. An electret is an object possessing a permanent electrostatic charge, making it rather like the electrical analog of the permanent magnet. Use of an electret to polarize the capacitor in a condenser microphone eliminates the need to supply a polarizing voltage for the capacitor. Since electret condenser microphones are used almost universally in cell phones, as well as other applications, they are probably the most popular type in current use. The diaphragm of electret condenser microphones is usually made of a thin sheet of flexible plastic which has had a metal coating deposited on it, a notably lightweight arrangement that permits good sensitivity and frequency response. Such microphones also usually exhibit good sample-to-sample consistency, making them practical for multi-microphone arrays.

A more recent variant of the condenser microphone is the silicon microphone. This is a microscopic condenser microphone constructed using micro electromechanical system (MEMS) technology. The required electrostatic polarizing field is provided by a charge pump, and the diaphragm can be free floating, being held in place by electrostatic attraction.

In recent years, the motion of a diaphragm has been measured using optical means, including laser interferometry, which may in time become a practical alternative to the capacitor microphone.

Microphone Dynamic Range

The noise level of a microphone is limited by Brownian motion of air molecules impinging on the diaphragm. Of course, it can be further limited by electrical noise of the amplification which follows it, but even discounting those effects, the so-called self-noise of a microphone is likely to be 10–20 dB higher than the threshold of hearing [18.30]. Lower self-noise is possible with larger diaphragms, say on the order of one inch, but this comes at a price of making the microphone more directional at high frequencies.

The loud signal limit of a microphone is often set by the maximum signal capacity of the associated preamplifier. Using large rail voltages can be of some aid in this regard. Eventually the diaphragm will reach the limits of its travel, and experience soft clipping. There has been some work on using force feedback to linearize the characteristics of the diaphragm [18.31, 32]. With a little effort and care, it is possible to make microphones that can handle high-level signals about as well as the human ear can.

Finally it should be noted that there has been some work on creating a microphone that does not require a diaphragm. Blondell and Chambers did some work on a flame microphone in 1902 and 1910 [18.23]. More recent efforts have used lasers to detect index of refraction changes due to acoustic pressure waves [18.33].

Microphone Pickup Pattern

Interacting with a real, three-dimensional sound field, a microphone is an audio device inextricably tied to spatial considerations. Immersed in the sound field, the microphone responds to the sound at essentially just one point, so its output represents a spatial sample of the total sonic event. This is not to say that it necessarily responds equally to sounds arriving from all directions; on the contrary, most microphones exhibit some directional bias, whether or not intended, and it is probably not possible to specify a single universally preferred pickup pattern.

The simplest arrangement is to put a sealed chamber behind the diaphragm, allowing only a slow leak for stabilization of quiescent atmospheric pressure. With short-term pressure behind the diaphragm fixed, the microphone responds to the sound pressure at the front of the diaphragm. Since pressure is a spatial scalar, it might be expected that a pressure microphone might respond equally to sound from all directions, and indeed this might be the case if the diaphragm had zero dimension. As it is, a finite-dimension diaphragm will spatially integrate the sound pressure across its surface. This will tend to increase output for normal-incident waves and waves with long wavelength, but will have less boosting effect for random incidence sounds, such as Brownian

Amazon EX1014 Page 767

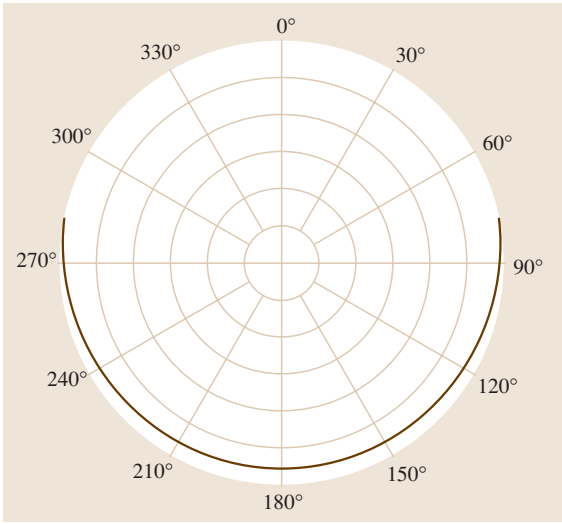


Fig. 18.13 *Omnidirectional* microphone pickup pattern. Note the slight attenuation of sounds arriving from the rear (After [18.35])

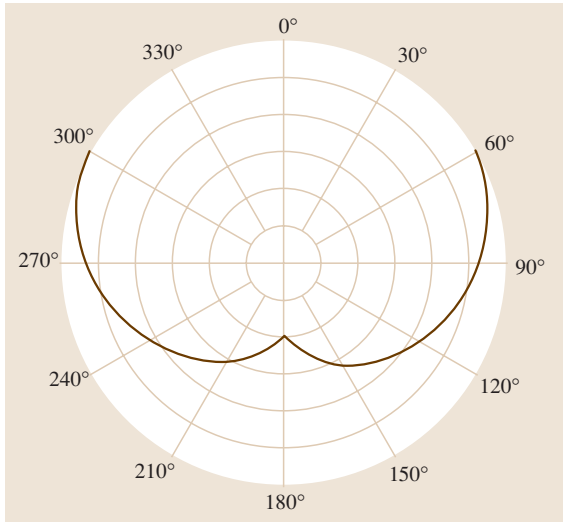


Fig. 18.15 Cardioid microphone pickup pattern, obtained by combining omnidirectional and bidirectional pickup patterns (After [18.35])

noise, and may reduce response to sounds with oblique incidence and wavelengths shorter than twice the diameter of the diaphragm. The resulting pickup pattern may therefore become directional at high frequencies, while remaining omnidirectional at lower frequencies. To remain omnidirectional up to 20 kHz, a diaphragm will have to have a maximum diameter of half the wavelength at the upper frequency limit, or about 8 mm. Whether by accident or design, this happens to be the approximate diameter of the human eardrum [18.34]. The fact that the eardrum is so small, and is coupled to the structures of the inner ear, and yet the ear still manages to exhibit a maximum sensitivity of 0 dB SPL or better has to be considered impressive performance, especially considering that quiet sounds just fade into perceived silence, rather than some inherent background perceptual noise.

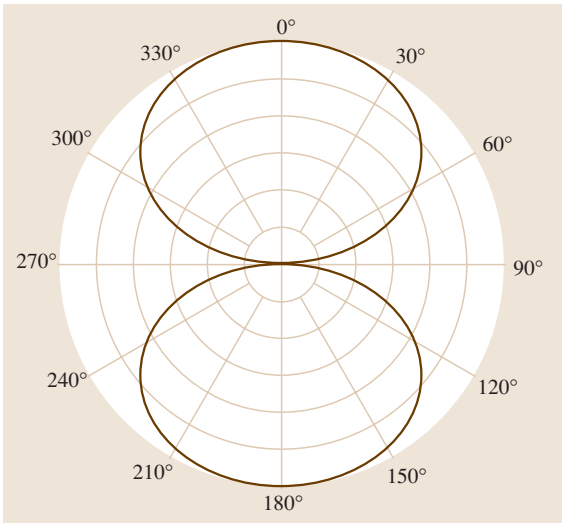


Fig. 18.14 Bidirectional microphone pickup pattern, typical of an open-back ribbon microphone (After [18.35])

More elaborate microphone pickup patterns (Figs. 18.13–18.15) can be achieved by increasing the number and complexity of acoustic paths to the diaphragm. For example, simply opening the back surface of the diaphragm to the atmosphere, the arrangement used in the classic open ribbon microphone, results in a microphone sensitive to the difference between front and back pressure, so sounds arriving from the side produce no response, and the pickup pattern usually resembles that of Fig. 18.8, albeit with the front arrival signal of opposite phase to the rear arrival signal.

If a pickup pattern as in Fig. 18.8 is summed appropriately with an omnidirectional pickup pattern, the rear arrival sound can be made to cancel, producing a cardioid front-directional pickup pattern.

With more elaborate channeling of the audio reading the diaphragm, more specialized or directional pickup patterns may be obtained.

Greater flexibility in the choice of pickup pattern can be obtained by using arrays of microphones, typically arranged in fixed physical orientations, where the

Amazon EX1014 Page 768

effective pickup pattern of the array is dependent on how the individual microphone signals are electrically summed. By recording the individual signals, the array can be effectively focused and oriented after the fact by postprocessing the recorded signals. The large-volume low-cost manufacture of certain types of microphones has made this approach especially attractive.

18.4.2 Records and Phonograph Cartridges

Although by now semi-obsolete, the phonograph cartridge has historically played a key role in sound reproduction, and is fairly close to a microphone in many aspects of operation. As with a microphone, the intent is to act as a transducer for acoustical vibration, in this case the vibration originating from a stylus following the undulations of a groove, rather than a diaphragm responding to vibrations in air.

Although early phonographs employed direct mechanical-to-acoustical transduction, the advent of electronic amplification made the electromechanical phono cartridge the default device for record playing.

Most of the methods of transduction used for microphones have been tried for phono cartridges as well, with piezoelectric and magnetic being the two most popular. Whereas most magnetic microphones have been of the moving-coil type, the magnetic cartridge design has employed both moving magnet and moving coil designs. Perhaps the use of a cantilever design of the stylus to reduce the effective mass of the coil or magnet seen by the stylus tip, coupled with the lack of a diaphragm, led to the preference of magnetic transduction for phonograph cartridges, while the use of capacitive transduction in condenser microphones has tended to be preferred in that transducer.

Despite the popularity of the grooved phonograph record for well over half a century, it has always been a difficult medium with which to achieve full fidelity.

For one thing, the stylus assembly is fraught with high-frequency resonance problems, much like the microphone, which have tended to limit the high-frequency response. In many cases, the high-frequency resonance has been underdamped, resulting in a serious peak in the mid-high-frequency response. The material used to mount the stylus in a way to secure it but still allow it to vibrate, often little bits of rubber, can sometimes degrade with age, impairing the characteristics of the cartridge. Then there is the need for the cartridge to drag a tonearm across the record with it.

In theory, the tonearm should provide a rock-stable platform for the cartridge at audio frequencies, so that only the stylus vibrates. Below the audio band, the tonearm should move as freely as possible, to follow the slowly spiraling groove and miscellaneous movements induced by the rotating disc, including horizontal eccentricity and vertical warp. In practice, this behavior cannot be achieved perfectly (since a perfect *brick-wall* filter is impossible), but at least the cartridge and tonearm should be matched, and the low-frequency resonance peak resulting from the compliance of the stylus and the effective mass of the tonearm should be properly damped, to avoid a peak either in the low-frequency audio response, or in the subsonic region where a resonance could make the system overly sensitive to record warp. Even in the best arrangement of properly damped resonance centered an octave or so below the audio band, the system is only second order, so some low-frequency tonearm vibration and some subsonic stressing of the stylus from pulling the tonearm is inevitable [18.36], which can lead to distortion and uneven frequency response. To avoid the build up of acoustic traveling waves in the tonearm, it is desirable to have its construction and material chosen to damp out such vibrations; not all tonearms have done this well.

Another problem that has hampered record playback, although not strictly the province of the phono cartridge, is that the playback stylus usually does not exactly follow the same path and orientation as the cutting stylus. This is because discs are cut on a cutting lathe, in which the cutter slowly traverses a straight-line path from the outer grooves to the inner grooves. The playback cartridge, in turn, is usually mounted on a pivoted arm, so its path is an arc. (The longer the tonearm, the smaller the disparity between record and playback orientations.) In order to minimize the path mismatch, a right dogleg bend has often been incorporated into the shape of the tonearm. The cartridge has to be carefully mounted on the tonearm so that the stylus is a specified distance from the pivot if this feature is to be properly exploited. Unfortunately, the bent tonearm traversing an arc results in a steady-state force from friction between the moving record groove and stylus tip, one which does not pull straight down the length of the tonearm, but instead generates a sideways force, called the skating force, which pulls the stylus tip against one or the other of the groove walls, which is another potential source of distortion. The skating force varies with groove speed and stylus tip size and shape, among other things, making it challenging to fully compensate.

At various times straight-line tracking tonearms have been produced, usually with the back of the tonearm mounted via bearings to a platform intended to move

Amazon EX1014 Page 769

in a straight line to follow the front of the arm (and the stylus). Since the stylus cannot be expected to drag the rear platform along with it, some sort of servo is used to insure that the platform follows the slow movements of the stylus/cartridge combination. Although this should mitigate skating-force problems, assuming the cartridge is mounted precisely, it still leaves the other difficulties of disc playback.

Another issue with phono cartridges is the shape of the stylus tip. The cutting stylus is chisel-shaped, with sharp corners, so it is capable of inscribing very fine undulations. But the traditional stylus tip shape is round, so the points of contact with the groove walls vary with the curvature of the groove. A round stylus will therefore follow a slightly different path than the cutting stylus, another source of distortion. In the 1960s, elliptically shaped styli became popular to reduce the tracing error, by more closely approximating the shape of the cutting stylus. The smaller stylus tip radius of curvature at the point of contact, compared to a spherical stylus tip, should allow the stylus to more accurately track the motion of the cutting stylus, but results in higher localized pressure, which can increase groove wear unless low tracking pressure is employed. Even so, the precision with which a given playback stylus follows the motions of the cutting stylus is unlikely to be sufficient to avoid some alteration in sound quality.

Around the time elliptical styli were becoming popular, RCA attempted to improve tracing performance by pre-distorting the cutter signal, anticipating playback with a round stylus. The system was not universally welcomed [18.37], in part because it arguably impaired performance with elliptical styli, and was eventually dropped.

Ultimately it has to be accepted that there will be some, hopefully slight, difference between the motion of the cutting stylus and the motion of the playback stylus, and that there are some signals that the playback stylus cannot hope to track properly, such as a constant-amplitude-cut square wave. The best one can hope for is that the deviation will be small and gentle, and that the resulting distortion will be mostly low-order THD. Even so, distortion figures for records were not uncommonly in excess of 10%, especially for the slower inner grooves. The technique of stereo recording on discs, pioneered by Blumlein in 1931 and introduced commercially in 1957, further complicated the task of the playback stylus, as it then had to track both horizontal and vertical undulations simultaneously.

Getting adequate dynamic range from records was a tricky process, fraught with compromises. For one thing, the maximum signal level used tended to be traded off against total playing time. Large-amplitude grooves took up more real estate than quiet grooves, which led to the use of variable-pitch, look-ahead record cutters that could dynamically adjust groove spacing in response to groove amplitude. Such cutters functioned by increasing the spacing one revolution before the arrival of a loud sound, and maintaining the spacing for the revolution that followed. Further complicating the amplitude/playing-time conundrum is the desire on the part of some recording engineers to minimize use of the inner groove area, to minimize inner groove distortion and bandwidth limitations.

Maximum amplitude was also limited by considerations of what the playback styli and cartridges of the consumers could accommodate, which could lead to lowest-common-denominator compromises in performance. To better match tracking ability and noise floor to the average statistics of music spectra, the Recording Industry Association of America (RIAA) standard equalization curve was adopted in 1954 that progressively reduced low frequencies and boosted high frequencies on recordings, applying inverse compensation during playback. This, along with the characteristics of the stylus, record surface and groove speed, made the overload characteristic of records strong functions of both frequency and playing time, so the precise specification of overload characteristic of records is problematical at best. To be sure, there are specified suggestions for maximum cutting amplitude and velocity, but given the commercial realities, at least with pop music, that louder all too often equates with better, the suggested limits have not always been observed.

The noise level of a record is dependent on the smoothness with which a groove can be cut, which in turn depends on the cutting technique and the material used. Although wax and shellac were used in the early years of record production, more modern practice has been to use lacquer deposited on an aluminum disc base. In conjunction with the hot-stylus cutting technique introduced in 1951 [18.9], a lacquer master disc is capable of a dynamic range approaching 70 dB, although in practice a somewhat lower figure is sometimes encountered for the end product on a production basis. Most of the steady-state noise encountered with records is concentrated at mid to low frequencies. In pursuit of the goal of fitting a performance to the dynamic-range characteristics and limitations of grooved discs, it has traditionally been common practice to employ a combination of compression and limiting to the signal in the course of cutting a master disc. There have been some

Amazon EX1014 Page 770

attempts to apply complementary compression and expansion to record audio to improve the system dynamic range, notably discs produced by dbx Inc. in the 1970s.

Successful record playback can also run afoul of a problem known as acoustic feedback, if sound from the loudspeakers causes the disc itself to vibrate as a diaphragm. Solutions include the use of an inert turntable mat, remote positioning of the turntable, and limiting the playback level.

Of course, dust, dirt, scratches and wear will tend not only to degrade the recorded signal but also to raise the noise level, including the possible introduction of clicks and pops.

Given the litany of compromises and limitations in disc recording and playback, the level of performance that has been achieved with this medium is laudable. Although largely eclipsed by digital audio, some refinement of grooved record playback continues, including efforts to achieve the holy grail of disc playback, contactless playback via light or laser beam, notably by Fadeyev and Haber at Lawrence Berkeley National Laboratory [18.38]. The availability of advanced DSP has allowed after-the-fact improvements in noise reduction and distortion reduction that would have been impossible in an earlier age. It is tantalizing to consider the performance that might have been possible with this medium if such processing had been available and incorporated into the basic architecture of the disc recording system.

18.4.3 Loudspeakers

Loudspeakers and headphones provide complementary transduction to microphones, namely converting electrical audio energy to acoustical energy. Many of the principles and notions regarding microphones can be applied to loudspeakers either directly or by, in a sense, turning the ideas upside down.

In a manner reminiscent of microphones, few loudspeakers convert directly between electrical and acoustical energy. Instead, the electrical energy is converted to mechanical energy, specifically the movement of a diaphragm, often cone-shaped or dome-shaped, which then radiates acoustical energy, sometimes aided by an acoustical coupling horn or lens.

Indeed, many of the same transduction methods that have been used in microphones have at least been tried for loudspeakers including magnetic (also known as *dynamic*), piezoelectric, and condenser.

The genesis of loudspeaker-like transducer actually predates Bell's invention of the telephone by a couple of years when, in 1874, Ernst W. Siemens described a dynamic moving-coil transducer, although he does not seem to have initially done much with it [18.39].

Bell's telephone, patented in 1876, also used a dynamic receiver. With a magnetic armature mechanically coupled to a diaphragm secured at its edges, it was probably closer in spirit to a headphone than a loudspeaker.

There followed a series of refinements that progressively brought the output transducer closer to the modern concept of a loudspeaker. In 1898, Oliver Lodge patented a spacer to maintain the air gaps in the magnetic circuit. Three years later, John Stroh came up with the notion of attaching the cone to the frame via a flexible, corrugated surround that improved the linear travel of the cone. Proper, flexible centering of the rear of the cone in the magnetic gap was provided by the *spider*, developed by Anton Pollak in 1909.

Many of these early loudspeakers used an acoustic horn of some sort to project the loudspeaker output into the room. The horn improves efficiency, partly by improving the acoustic coupling from the loudspeaker to the room, and partly by collimating the sound into a beam directed at the listener. Credit for the first true direct radiator loudspeaker design to include all the elements of what can be regarded as the modern-day loudspeaker is generally accorded to Rice and Kellogg in 1925 [18.40]. Their design took explicit account of the baffle on which the driver was mounted in the overall acoustic design.

There were of course, countless follow-on refinements to the basic electrodynamic driver, a process that continues to present day. Other methods of transduction were also explored. In 1929, E. W. Kellogg filed a patent for an electrostatic loudspeaker, consisting of panels not unlike large condenser microphones. The design has not been as widely accepted as the dynamic driver, in part due to some limitations in maximum output. There have also been piezoelectric drivers, but the limited travel of the piezoelectric element has limited their use to mid-range drivers, tweeters, and earphones.

So, as the Rice/Kellogg baffle indicated, with the basic dynamic loudspeaker configuration established, much of the attention focused on refining the loudspeaker/cabinet combination as a system.

The output ultimately obtained from a raw loudspeaker driver is profoundly affected by the surrounding baffle or cabinetry. In isolation, a loudspeaker driver produces two acoustical outputs: one from the front of the diaphragm or cone, and the second from the back. Unfortunately, the two outputs are out of phase. Were they generated at the same place and time, they would can-

Amazon EX1014 Page 771

cel; the loudspeaker would produce nothing but heat. So part of the goal in designing a complete loudspeaker system is to keep those two outputs from annihilating each other.

At frequencies above the point where the wavelength is comparable to the size of the speaker cone, the radiation will tend to be a directional beam, and the front and back waves will tend to naturally go their own separate ways. At lower frequencies, the front and back radiation will tend to be more omnidirectional, and will therefore tend to cancel. This can be avoided by requiring that the size of the loudspeaker be at least as large as the wavelength of the lowest frequency of interest. For a low-frequency limit of 20 Hz, this would correspond to a diameter of at least 15 m, which at the very least would make for rather unwieldy boom boxes.

Happily, it is not necessary to rely on the loudspeaker itself to avoid low-frequency cancellation. Instead, a smaller driver can be mounted on a baffle, and the baffle can be used to prevent interaction of the front and back radiation. Indeed, a common concept in loudspeaker system theory is that of the *infinite baffle*, which precludes cancellation at any frequency, at the cost of being physically impossible. Even assuming a willingness to forego response below 20 Hz, a baffle of at least 15 m would still be required, which is not much more practical than an unadorned 15 m woofer. This leads to the notion of putting the loudspeaker in a box, or loudspeaker cabinet. (A speaker mounted in a car door or trunk lid is operating in an approximate infinite baffle, although the car interior can be considered a semi-sealed cabinet.)

The simplest arrangement is to make the box completely sealed, so that the rear radiation is trapped within the box, and cannot emerge to cancel the front radiation. Of course, this requires making the box quite rigid, else the rear speaker radiation will simply be conducted through the walls of the box. One problem with sealed-box systems is that the springiness of the air in the cabinet will decrease the natural compliance of the loudspeaker, so that a driver with a respectably low free-air resonance may have a much higher resonance mounted in a sealed box, with correspondingly reduced low-frequency response.

This issue was addressed in 1954 by Edgar Villchur, in a manner which in hindsight seems remarkably straightforward, namely increasing the compliance of the speaker to obtain the desired performance when it was mounted in a sealed box. This approach, referred to as an acoustic suspension design, made the cabinet an inherent part of the design, and is still in use today.

The alternative to the sealed-box loudspeaker enclosure is, logically enough, the vented box, in which at least a portion of the back radiation of the driver is allowed to emerge from the cabinet, often delayed or phase-shifted to better align it with the front radiation. One of the early examples of this approach was the bass reflex design by Albert Thuras of Bell Laboratories, in 1930. Other vented designs have used elaborate internal tubing, internal ductwork (such as the Klipsch folded horns), and or provided mass loading of the exit port via a diaphragm (the *passive radiator*). The goal is to improve efficiency, partially by relieving some of the back pressure behind the driver, but mostly by adding some of the back radiation to the front radiation in a nondestructive fashion, to achieve some increase in output, typically on the order of 3 dB.

More recently, hybrid enclosures have been developed, sometimes called bandpass enclosures, in which a vented or sealed box is used behind the driver and a vented box is used in front of the driver, to provide improved low-frequency efficiency over a somewhat narrow range of frequencies [18.41]. This design is popular in some smaller low-frequency speakers.

Regardless of its other characteristics, a bare loudspeaker cabinet is likely to impart a variety of spurious resonances to the response from sound waves bouncing helter-skelter around the inside of the cabinet. For this reason, loudspeaker cabinets are often lined or stuffed with sound-absorbing material. This alters the overall response as well as damping the resonances, and must be accounted for in the overall design.

Much of the modern modeling of loudspeaker systems is based on the work of Thiele and Small in the 1970s, who established the basic parameters governing the performance of loudspeaker systems, and developed accurate mathematical models incorporating those parameters, allowing rapid and precise mathematical prototyping [18.42].

Still, the implications of achievable, cost-effective drivers and practical cabinet size is that there is precious little that can be done to get usable bass response down to 20 Hz from a small loudspeaker system. It is possible to apply low-level equalization to the signal upstream of the loudspeaker, but this can increase the cost and the power requirements, and will not be effective unless the driver can handle increased signal level without serious distortion. The closed-box design is considered one of the better choices for the use of equalization, as its response roll-off below primary resonance is only 12 dB per octave, which is gentler than most vented systems.

Amazon EX1014 Page 772

Most of the issues relating to speaker cabinetry pertain to low frequencies. Midrange and high-frequency drivers may also have their backs sealed, but the wavelengths are shorter, and the required resonant frequencies are a lot easier to attain.

The notion of using multiple drivers in a loudspeaker system seems to have originated in 1931 with a two-way system designed by Frederick of Bell Laboratories [18.43]. A three-way system followed within two years. It is simply much easier to design a system to have flat frequency response with multiple drivers, optimized to handle a subset of the audio band, than to try to accomplish the task with a single driver.

Part of the motivation for using multiple drivers is that the loudspeaker, like the microphone, is an inherently spatial entity. Sound radiates from a loudspeaker in all directions, depending on angle and frequency. In an anechoic environment, with a listener situated on the main axis of the loudspeaker, the response on that axis is well defined, and the only one that really matters. In a real room, however, the multidirectional acoustical output of the loudspeaker will reflect off walls and other surfaces, generally producing a complicated multiplicity of arrivals at the listener's ears. There is no easy way to characterize this process, or of the resulting perception, although in the 1970s, Roy Allison made significant strides in characterizing the loudspeaker–room interaction at low frequencies [18.44]. However, it certainly establishes the importance of the loudspeaker spatial response at all angles, and suggests the desirability of a loudspeaker having a consistent radiated frequency response (radiation pattern) at any angle.

This is another area in which multiple drivers can be beneficial. Again, the output of a driver will begin to get directional at a frequency corresponding to a wavelength about twice the diameter of the driver. Assuming consistently wide dispersion is desired, a 30 cm woofer will be useful up to about 500 Hz, a 7.5 cm midrange continuing to 2 kHz and a 2.5 cm tweeter maintaining wide dispersion to at least 6 kHz. These figures may be conservative, in part because controlled decoupling of the speaker cone can result in a progressively smaller radiation area at successively higher frequencies, decreasing the effective diameter and broadening the resulting dispersion. Other approaches to maintaining a desired radiation pattern across frequency include the use of acoustic lenses and phased arrays of multiple drivers within a given frequency range. It is rare for a speaker to have truly constant directionality across its entire pass-band, in part from that fact that most are at least somewhat directional at mid and high frequencies, and, because of the long wavelengths involved, almost unavoidably omnidirectional at low frequencies. (Such long wavelengths may be difficult to localize, enabling the use of satellite systems with only one woofer to support two or more channels.)

The use of two or more frequency-selective drivers generally imposes the need for a crossover network to apportion the signal on the basis of frequency. Not only does this avoid wasting power, but it can be absolutely necessary to avoid overdriving or burning-out a driver. Some care is generally required in the transition frequency region to maintain flat response, and although there are well-regarded canonical topologies, such as the Linkwitz–Riley crossover developed in 1976 [18.45], real-world deviations from ideal driver behavior may engender the use of recursive approximation techniques to optimize system response fully.

The combination of crossover networks, drivers, and room loading will often result in a loudspeaker having an electrical input impedance characteristic that is strongly frequency dependent, and which poses a challenge for a power amplifier.

As is no doubt evident, a loudspeaker system and its associated radiation pattern is a fiendishly complicated device that does not readily lend itself to characterization by traditional audio metrics or specifications. In consideration of such specifications, it is difficult to make definitive statements that fully characterize the sound of a speaker.

The bandwidth of most speakers may approach the ear's 20–20 kHz bandwidth, but few speakers make it all the way, particularly at low frequencies, and bandwidths on the order of, say, 70–14 kHz are far more common, at least in consumer equipment.

The deviation of the frequency response is difficult to talk about meaningfully, but it seems desirable for speakers to be within 1 dB of flat response across their useful bandwidth, at least on axis, and better speakers can approach or achieve this, although at low frequencies the room exerts enough effect as to likely require tuning or experimentation with speaker position to maintain a flat response. The response at other angles should probably be smooth and well controlled, but the exact optimal specification has not been established. Perhaps the ideal speaker would adjust its radiation pattern to match the instruments being reproduced. (But then what is the preferred radiation pattern of a purely electronic instrument, such as a synthesizer?) Although it is necessary to measure the response of a loudspeaker in a room in order to accurately gauge the low-frequency response, a direct narrow-band or sine-wave measure-

Amazon EX1014 Page 773

ment of the system response in a room at middle and high frequencies is likely to yield a useless, mad scribble of a response, owing to the complex interactions of the echoes at the point of measurement. Fortunately, the perceived response should be a lot smoother than what is measured, in part because the ear differentiates sound arrivals on the basis of direction. Measurement techniques that attempt to address this include the use of directional microphones, anechoic and semi-anechoic measurement, and time-gated response measurement.

Timing precision is another specification that is difficult to pin down with loudspeakers, again partly due to the effects of room echoes. In one respect, most speakers do pretty well in first-arrival timing, since the ear's temporal pre-masking resolution is on the order of 1–2 ms, corresponding to about 30–60 cm of path length, and speaker systems with all drivers mounted on a common front panel will generally be well within this tolerance for inter-driver path-length differences. Still, audible time disparities may result from primary resonances, particularly at low frequencies, and high-order crossovers.

In the area of distortion, loudspeakers also do pretty well, all things considered. At normal listening levels, a properly designed loudspeaker exhibits fairly low nonlinearity, with a fairly gentle limiting characteristic, leading to low IM distortion and fairly low THD, mostly third order, which is likely to be masked by most signals. Most properly operating loudspeakers do not produce drastic amounts of distortion until they reach the limits of their travel. To be sure, some loudspeakers can sound audibly harsh, but this is not necessarily a byproduct of distortion, it may be more related to an uneven response or radiation pattern. One bugaboo of many drivers is cone breakup, wherein the cone no longer acts as a piston. Although this can look unsightly in slow motion, it will not necessarily produce distortion unless there is nonlinearity present, and most speaker cones, being passive structures of paper, plastic, or metal in and of themselves, are not especially nonlinear. While such breakup is likely to affect the response and/or the radiation pattern of the speaker, it may not produce much distortion per se.

Although the basic dynamic loudspeaker has now been in use for over 70 years, and has been greatly refined in that time, it is still the subject of active research, in both design and evaluation. One of the holy grails yet to be achieved is, like the diaphragm-free microphone, the cone-less loudspeaker that more directly converts electrical energy to acoustical energy. Aside from eliminating some troubling mechanical elements and their associated resonances, this might improve the efficiency, which is currently typically around 1% for consumer loudspeakers.

Another long-sought goal is the projection loudspeaker, which can make a sound appear to emanate from a specified remote location. The have been some encouraging results in the use of ultrasonics [18.46] and large-scale speaker arrays [18.47] in this regard.

18.4.4 Amplifiers

Of all the common audio analog devices, electronic amplifiers of all types probably have the easiest time meeting the preferred specifications of a good-quality audio system. This in no small part derives from the fact that they generally contain no mechanical processing stages, and simply have to push a bunch of very lightweight electrons around. Properly designed, an amplifier should have little difficulty achieving a dynamic range approaching 120 dB, nor is it likely to deviate significantly from the preferred specifications for bandwidth, response deviation, phase/timing response, or distortion.

This is despite the fact that most amplification devices, mainly vacuum tubes, transistors, and field-effect transistors (FETs), are not themselves inherently very linear. In the years following de Forest's invention of the triode, considerable and rapid strides were made in not only the refinement of the devices themselves, but in circuit designs which optimized the linearity and response of the complete amplifiers. A major milestone was the invention in 1928 of the negative-feedback amplifier by *Black* of Bell Laboratories [18.9]. This arrangement compares the output signal with the input signal, and to the extent that they may tend to differ, generates an instantaneous correction signal. Successful application of negative feedback carries with it specific requirements on the performance of the raw amplifier, such as maximum phase shift, and some of these requirements were still being uncovered years after the initial development of negative feedback.

Several standard types of amplifiers are commonly found in audio use, each with specific requirements and design challenges, including preamplifiers, line amplifiers, power amplifiers, and radio frequency (RF) and intermediate frequency (IF) amplifiers for use in radio. Preamplifiers, for example, are required to handle the generally very small signal from transducers such as microphones, phonograph cartridges, tape heads, and optical photocells. Linearity is usually not a major problem, because of the small signals involved, but low noise and impedance matching to the transducer are often sig-

Amazon EX1014 Page 774

nificant design considerations. At the other end of the chain, power amplifiers usually have a fairly easy time with dynamic range, but high required power-handling capability and the need to drive sometimes ill-behaved loudspeaker loads make power amplifier design an art unto itself, with the 1948 Williamson amplifier being just one early example of a notable design [18.48].

The advent of the transistor, by Brattain, Bardeen, and Shockley of Bell Laboratories in 1947, sparked a revolution in active devices, and stands as one of the foremost inventions of the 20th century [18.49, 50]. Operational differences between tubes and transistors were sufficiently great that, to an extent, the discipline of amplifier design had to be reinvented from the ground up, and it was some time before solid-state amplifiers were accepted in some high-end audio systems.

In time, the development of analog integrated circuits gave rise to the multi-transistor operational amplifier, an amazingly versatile device that facilitated cost-effective analog processors of great sophistication and complexity. These days, entire analog audio subsystems can be implemented on monolithic chips with many transistors and associated components.

Special mention should be made of one particularly challenging amplifier variant, the voltage-controlled variable-gain amplifier (VCA). Such a device, basically an analog multiplier, is essential for exercising automatic control of signal level, a common building block in many audio signal processors. Where fixed-gain amplifiers depend on fixed, highly linear passive components such as resistors to maintain operating point and linearity, VCAs require linear active elements, like FETs. In the late 1960s, Barry Blesser designed a novel variable pulse width multiplier, and a few years later, David Blackmer of dbx, Inc. designed a wide-range variable-transconductance VCA, elements of which are still used in commercial VCAs [18.51]. The complexity and cost of VCAs limited their use somewhat, which is perhaps ironic in these days of digital audio, where DSP chips typically perform millions of multiplications per second.

18.4.5 Magnetic and Optical Media

Part of the motivation for the use of magnetic and optical recording on linear media (tape, film) undoubtedly came from the desire to avoid the wear inherent in the playback grooved records. Both media involve the use of narrow apertures to record and play audio striations perpendicular to the motion of a linear medium.

Magnetic recording got its start in 1898, with Valdemar Poulsen recording on steel wire, an especially notable accomplishment considering that the invention of the triode amplifier tube was still several years away. The use of magnetic tape originated with O'Neill's patent of paper tape coated with iron oxide, in 1926, followed by the introduction of plastic tape by BASF in 1935. A serious nonlinearity problem with magnetic tape was largely resolved in 1939 with the development of alternating-current (AC) recording bias. Considerable refinement was achieved during and following World War II, and consumer reel-to-reel tape recording became a reality in the 1950s. There followed a series of formats which expanded the performance envelope and consumer friendliness, most notably the compact cassette and the venerable eight-track cartridge.

Optical recording on film was first investigated around 1901 [18.9], five years before the vacuum tube, and started to receive significant research and development attention starting around 1915, with the work of Arnold and others at Bell Laboratories, followed by a number of efforts by RCA and other groups. Lee De Forest may have come up with the first viable optical sound system, *Phonofilm*, in 1922 [18.52]. By 1928, synchronized optical sound on film was used commercially on Disney's *Steamboat Willie* cartoon.

Both of these formats rely on some sort of device with a narrow aperture to record and playback, and therefore both face a tradeoff between high-frequency response, corresponding to the smallest resolvable feature, and media speed/playback time.

Although there may be no vibrating mechanical elements in either the magnetic system or optical playback, both media exhibit difficulty reaching the full audible bandwidth, or in maintaining ruler-flat response, although, properly designed and adjusted, they can come fairly close. Beyond that, both media bear some similarities to modern record-based reproduction, including the use of fixed equalization, modest, gentle distortion at typical operating levels, a typical dynamic range on the order of 50–70 dB, and both absolute and dynamic speed error sensitivity.

The viability of these formats was significantly enhanced by the invention in 1965 of analog noise reduction by Ray Dolby. Noise-reduction systems usually consist of a compressor/encoder and an expander/decoder. The compressor boosts low-level signals to keep them above the noise floor of the channel, while the expander restores the signals to their proper level, reducing the noise in the process. The Dolby A-type noise-reduction system and follow-on systems exploited a range of solid-state circuit tools, including active filters, precision rectifiers, and lin-

Amazon EX1014 Page 775

ear FET-based VCAs, to implement a comprehensive solution to the problems faced by noise-reduction systems. The frequency-selective nature of masking by the human ear was addressed by splitting the signal processing into multiple bands, and the problem of potential overload associated with the sudden onset of loud signals was resolved by the use of a novel dual-path limiter. Other well-known NR systems used wide-range bandpass compressors (Telcom C-4), wideband compressors with preemphasis (dbx I and II), sliding shelf filters (Dolby B), cascaded sliding shelf filters (Dolby C), cascaded wideband and variable slope compressors (dbx/MTS TV compressor), and cascaded combined sliding shelf and bandpass compressors (Dolby SR and S). Similar principles would subsequently be adapted for use in digital low-bitrate coders.

18.4.6 Radio

While the medium of analog radio has its own share of unique attributes, it shares with other analog media many of the same sorts of characteristics and limitations, including somewhat limited frequency response and dynamic range, which is partly associated with bandwidth conservation.

Commercial amplitude modulated (AM) radio broadcast kicked off in 1921 when KDKA went on air in Pittsburgh [18.9]. The use of dual-sideband modulation made for simpler receivers at the cost of channel spacing being twice the highest transmitted audio frequency. Although in theory AM radio was capable of supporting wide audio bandwidth, a channel spacing of just 10 kHz was chosen to conserve spectrum space, limiting audio bandwidth to something less than 5 kHz in practice.

Wide-band frequency modulated (FM) radio was developed by Edwin Armstrong in 1933, motivated in part by a desire to reduce the static that was common to AM radio, with commercial broadcast originating in 1941 [18.9, 53]. The dual sidebands of standard FM extend quite a bit farther than the maximum audio frequency supported, with channel spacing of 200 kHz for a typical audio bandwidth of 15 kHz. Dynamic range in excess of 60 dB is possible with mono FM, although this was compromised somewhat by the choice (in the US at least) made by the Federal Communications Commission (FCC) of the stereo multiplex system in 1961 [18.9].

18.5 Digital Audio

The advent of digital audio, perhaps best signified by the introduction of the audio compact disc in 1982 by Sony and Philips, completely revolutionized the world of audio engineering, in effect completely rewriting the rule book on what was readily achievable in mainstream audio systems, not to mention how signals were processed.

The basic notion of digital audio is discreteness. Instead of specifying a signal as a continuous function of time and amplitude, it is expressed as a rapid series of sample values at discrete time intervals, often with integer or finite-precision values. The signal becomes, in effect, a completely specified list of numbers, and as long as no explicit errors are made, can be copied an arbitrary number of generations without any loss or alteration whatsoever. (Error-correction schemes commonly help guard against the occasional error slipping through.)

The mathematical foundations of discrete sampling of continuous functions go back surprisingly far, with at least one claim on behalf of Archimedes, around 250 BC [18.54, 55]. Additional support for sampling theory was provided by *Fourier* in 1822 [18.56, 57] and *Cauchy* in 1841 [18.58, 59]. Modern sampling theory is usually credited to the combined work of *Whittaker* (1915) [18.60], *Nyquist* (1928), *Kotelnikov* (1933), and *Shannon* (1949) [18.61, 62], who provided a rigorous proof [18.63].

The key point of the sampling theorem is that a continuous band-limited signal can be discretely sampled without loss of information, and the original continuous signal subsequently recovered from the sampled version, as long as the sampling frequency is higher than twice the signal bandwidth.

In order to exploit the sampling theorem and digital audio practically, quite a few other pieces of the puzzle had to be supplied. George Boole contributed the binary number system in 1854, which would provide the numerical and logical basis of modern digital processors. Alec Reeves used Boole's binary numbers to represent audio samples, creating the pulse code modulation (PCM) format, in 1937. By the late 1960s, prototype digital audio recorders were operating at companies like Sony, NHK, and Philips [18.58, 59], leading to the 1982 debut of the audio CD.

Amazon EX1014 Page 776

Some appreciation of the performance of digital audio can be had by considering the traditional audio specifications that analog systems have for so long tried to fully embody.

Bandwidth. Traditionally an uphill battle fighting mechanical resonances and slit-loss effects, successfully meeting a bandwidth specification in the digital domain is simply a question of selecting a high enough sampling rate. Similar principles apply to maintaining the precision of a frequency response specification, usually a flat response: without parasitic or extraneous influences, the response will be ruler practically-flat by default, although it can otherwise be explicitly and predictably altered by application of appropriately designed DSP filters and equalizers.

Distortion. With signal processing implemented as a series of digital additions and multiplications, which are presumably accurate to the least significant bit, nonlinearity and distortion in the digital domain is absent unless, like deviation from flat response, it is explicitly introduced into the processing, to which end it can usually be precisely applied and controlled.

Dynamic range and signal-to-noise ratio. The dynamic range is usually expressed as the logarithmic difference between the maximal signal level and the noise level. By default, the maximal signal capacity of a PCM DSP system as a function of frequency is flat. The noise level is once again a design choice, ergo so is the net dynamic range. In somewhat simplified terms, the noise level of a digital signal process drops by a factor of two (6 dB) for each additional digital bit of precision employed in the representation of the sample values. So, maintaining a desired dynamic range is fundamentally a design choice to employ a sufficient number of bits of precision.

Speed. Once a signal is in the digital domain, there is of course no error or variation in speed unless it is explicitly imposed with appropriate processing.

In short, the attainment of performance that in the analog domain might require considerable design skill, and even then might simply be impractical or impossible, becomes in the digital domain a matter of design choices.

Of course, since most audio signals begin and end as analog vibrations in air (and, so far, correspondingly analog electrical signals associated with input and output transducers), there must be means to convert between analog and digital domains. This is the province of analog-to-digital converters (ADCs) and digital-to-analog converters (DACs). With so much of the overall performance of a digital audio system riding on the performance of these devices, their design has been the subject of intense activity since early experimental digital audio systems. Those early systems were fortunate to muster a precision of just 10-bits per sample, but that has slowly evolved over time to the point that current converters can operate at 24-bits per sample or better in a practical, cost-effective manner. Similarly, the various sources of subtle distortions that might otherwise limit ultimate converter performance have progressively been identified and, on the whole, reduced to the point of insignificance. And, where once speed accuracy of an audio system depended on the skill with which one turned a crank, or perhaps the vagaries of a wind-up clock motor and mechanical speed regulator, the sampling rate of a digital/audio converter is usually controlled by a high-precision electronic timing circuit, such as a quartz oscillator, reducing speed anomalies of all sorts to a level far below that of human perception.

Lest the case for digital audio appear at this point a bit overly rosy, it must be allowed that there are limitations, restrictions and pitfalls; and that it is perfectly possible to produce miserable sound via digital means.

For one thing, there is the matter of the Nyquist frequency, defined as half the sampling frequency and otherwise the highest frequency that can be represented at a particular sampling frequency. In some respects, the Nyquist frequency of a digital audio system corresponds to infinite frequency of an analog system. This introduces certain response-warping effects, such that it may be difficult or impossible to replicate a given analog system response in the digital domain. (Example: the phase shift at the Nyquist frequency can only be 0° or 180°.)

An attempt to represent a frequency greater than the Nyquist frequency in the digital domain will result in a different frequency being generated within the band, usually by reflecting the original frequency value back through the Nyquist frequency, a form of intermodulation distortion. Unless the goal is to produce weird spacey effects, this process is usually not highly desirable, and a low-pass filter is commonly placed ahead of an ADC to prevent that occurrence.

Then there is the issue of performance levels being matters of system architectural choices, rather than the result of possibly elaborate or laborious designs. Such

Amazon EX1014 Page 777

design choices are not always free, in that they may be tradeoffs with either other system resources, or simply cost more monetarily, which may therefore still have the result of limiting overall performance to something less than idyllic near-perfection.

For example, the dynamic range of the human ear was cited earlier to be on the order of 120 dB or more. If the digital representation of an audio signal yields an approximate dynamic range of 6 dB for each bit used, it would follow that the logical word size to use for digital audio samples would be at least 120 dB/6 dB per bit = 20 bits. Yet the standard chosen for compact discs, as well as many other current generation digital devices, is just 16-bits per sample, corresponding to a best-case dynamic range of only 96 dB. This decision was predicated in part on the performance limitations of analog/digital converters at the time the CD specification was cast, and further from a desire to provide a playing time of 74 min on a 12.7 cm disc. Had a larger word size been chosen, either the disc would have had to be larger, or the playing time reduced.

Does this result in audible impairment of the sound? It can, but it will not necessarily be the case, so it will be evident some hopefully small percentage of the time. For example, there should be no audible alteration if the maximum playback level does not exceed 96 dB SPL. Nor should there be a problem if the playback level is higher, but the room noise is high enough to mask the digital noise. Since the ear is only maximally sensitive in the upper-middle part of the audio band, the elevated thresholds towards the band edges provide additional noise margin. And the presence of signal components will tend to mask digital noise that might otherwise have been heard. So, consistent with practical experience, 16 bit PCM digital audio is free of serious imperfections for most listeners and most source material, most of the time.

18.5.1 Digital Signal Processing

With the practice of representing audio signals in the digital domain as PCM samples comes the need to perform signal processing in the digital domain. To some extent, this required a reinvention of signal processing in terms of discrete numerical sample values, but since the foundations of signal processing were mathematical to begin with, in some respects there was a more direct connection between theory and practice.

The most common signal processing operations – amplification, filtering, and equalization – are mostly implemented with the basic four mathematical functions: addition, subtraction, multiplication, and, perhaps to a lesser extent, division.

More elaborate signal processing functions, such as rectification or level derivation, may involve geometric or transcendental functions such as square roots, cosine, or logarithms, virtually all of which are easier to implement, more direct, more wide-range, and more stable as digital entities than analog alternatives. Even some seemingly simple functions, such as a pure delay, can be challenging to implement well as analog devices, but become more-or-less trivial with DSP.

Clearly, the requirement of audio DSP to perform large numbers of numerical calculations and the ability to perform such calculations with modern digital hardware, including personal computers and specialized DSP chips, whether by accident or design, has been auspiciously symbiotic. Large, elaborate analog breadboards of spaghetti wiring have been replaced with even more elaborate programs of (sometimes) spaghetti C code.

The ability to create and combine a virtually limitless number of new processing functions by simply typing in the code for them, using a computer language with full mathematical support, is an understandably powerful tool to extend signal processing, a little like a bottomless Tinker Toy. The effect is amplified by the ability to nest functions to an almost arbitrary extent, allowing the creation of labyrinthine tree-structured processes. And whereas every block of an analog circuit must be physically implemented, DSP programs usually get by with just one physical copy of each routine, regardless of how many times the function may appear in the block diagram.

With all this processing power available, one may inquire as to what has been applied. Certainly, one function category that has *not* been strongly needed has been aids to correct deficiencies inherent in the recording system. If a clean recording is obtained to begin with, there is often little need to clean it up further. Digital signal processors have, however, been very effective in cleaning up and correcting the imperfections of older, analog recordings, and have been used for noise suppression, speed correction, and even distortion reduction.

The additional processing power afforded by DSP processors has enabled re-engineering of most traditional audio processors, such as equalizers, with greater complexity and precision. The availability of a simple, cheap delay mechanism with equalization has allowed extremely elaborate and faithful reverberators to be employed in both professional and consumer applications. This is one area for which there was arguably never a truly effective analog electronic solution. The best

Amazon EX1014 Page 778

pre-digital reverberators were electromechanical, using transducers attached to a metal reverberation plate.

In general, the availability of a cheap, low-noise, high-precision multiplication operation has, along with other DSP elements, made it much easier to implement elaborate dynamic systems, once limited by the relative scarcity and high cost of analog VCAs. This has also made it much easier to include specific nonlinearities, where appropriate, usually in the signal analysis of a processor. And the ability to combine multiple processing algorithms into a single DSP program has facilitated multifunctional systems implemented on a single DSP chip.

One of the key building blocks to be brought to bear in a variety of digital signal processors has been a digital model of some or all of the human auditory system. If a processor can be made to hear the same way as humans do, it can make processing decisions that reflect the truly audible elements of a signal, resulting in far more effective processing. A key component of such an approach has been the digital filter bank, essentially a bank of bandpass filters that, in most cases, cover the entire audio band. This component provides a cost-effective analog of the filter-bank-like processing performed by the basilar membrane of the inner ear. A major element of such processing has been the so-called fast Fourier transform (FFT), first devised in 1805 by *Gauss* [18.64–66], then rediscovered in 1965 by *Cooley* and *Tukey* [18.67, 68]. This one algorithm allows full-range filter banks of great complexity, typically implementing hundreds or thousands of bandpass filters, with dramatically less processing power than would otherwise be needed by direct implementation of such filters. The use of this technique does, however, carry with it some compromise, as the bandpass filters of an FFT filter bank all have the same bandwidth, measured in Hertz, and are linearly spaced in center frequency, whereas the filters of the basilar membrane are logarithmically arranged in both bandwidth and center frequency. This discrepancy is sometimes compensated by processing the FFT outputs in groups, or bands, whose width and center frequency are arranged in an approximately logarithmic manner.

The transform-based filter bank is used both as an analysis tool, to present signals in a perceptually relevant domain for subsequent processing, and as a component of the actual signal path. In the latter instance, signals are processed in the transform domain, usually in overlapping blocks of samples, and ultimately restored to time-domain PCM values by performing running inverse transforms. This has elevated transform-based processing to the level of a new category of digital signal processing. In some situations, a variant transform of some sort may be used to provide, for example, logarithmic frequency selectivity, but in many cases such transforms are still based on a core FFT.

A final note on numerical precision: virtually any digital processor will employ finite-word-length symbols for numerical values, be they integers or floating-point numbers. This gives rise to the generation of digital noise, resulting from the round-off error of each arithmetic operation. Since there may be many thousands of such performed in the course of a DSP algorithm, the numerical precision of the processor must be great enough that the accumulated noise remains insignificant. For a 120 dB dynamic range, corresponding to 20 bits of precision, one might prefer at least 24 bit processing (that is, either 24 bit integers or 24 bit mantissas of float values). The actual precision required will depend some on the actual operations performed, with subtraction (or addition of opposite-signed values) especially troublesome, since it can leave a very small residual, potentially exposing the digital noise. As it happens, this is an issue in performing the FFT. As an additional hedge against the buildup of digital noise, many DSP processors employ double-length registers, so accumulated quantities are only rounded back to standard precision when stored back to memory, usually after multiple multiply–accumulate operations.

18.5.2 Audio Coding

The commercial introduction of digital audio in the form of the compact disc brought with it the reincarnation of a tradeoff that had long been a source of compromise in analog audio, namely that between quality and capacity, or equivalently playing time, or bandwidth. Mechanical audio recordings (i.e. records) had started with cylinders that played for a minute or two, and had progressed to long-play records with about 20 minutes per side. Magnetic tape had used media running at speeds like 76 cm per second, and had progressed to cassette tapes with typical playback times of 45 minutes per side, usually aided with noise reduction. Broadcast channels were less flexible, but techniques had at least progressed to the practice of piggy-backing low-energy subcarriers onto to the main signal, e.g. to advance to stereo operation, sometimes accompanied by noise reduction, at least in the case of US multichannel television sound (MTS) audio. If the use of digital audio for recording and broadcast was to be adopted for media beyond the CD, the need to

Amazon EX1014 Page 779

reduce the number of bits required was apparent. This has given rise to one of the premier pursuits of digital development since that time: the use of audio coding to convey a signal faithfully using the smallest possible number of bits.

The data rate of an audio CD derives from the use of two channels, a sampling rate of 44 100 Hz, and 16 bit PCM samples, a total of 1.4 megabits per second. This is greater than can be accommodated on a standard 35 mm movie print, or a standard cassette tape, or an FM radio channel. (It is also more than will fit on a standard long-play record, but since that medium is not especially compatible with digital storage in the first place, it is a moot point.) Even the CD can only accommodate the two channels; a 5.1-channel PCM format would reduce the playing time to something under 25 min.

So the need to reduce the data rate without sacrificing quality was readily apparent, which motivated interest and activity in the development of low-bitrate audio coders, and made practical such applications as digital cinema soundtracks, portable music players, and satellite radio. It is perhaps not self-evident that the data rate can be reduced in the first place, but for starters, one could simply use smaller PCM words. This would, of course, raise the effective noise floor, which in turn could be addressed by adapting an analog noise-reduction system as an associated DSP algorithm. However, the resulting reduction in data rate would still be relatively modest, as the compression ratios of noise-reduction systems never exceeded a range of about 2–3:1, to avoid mistracking problems, and this is less data compression than is required by many coding applications.

Overall, two general classes of techniques have been exploited in digital audio coders, lossless and lossy, or perceptual, coding. They share in common the arrangement of an encoder and a decoder which, instead of just conveying PCM samples, employ a common vocabulary or protocol for describing and conveying the audio at an elevated level of abstraction. This protocol is sometimes referred to as a bit-stream syntax. Although the signal connecting encoder and decoder is indeed usually just a stream of bits, the meaning and significance of a given bit will depend on the context in which it appears. This represents an additional advantage of digital audio systems over analog systems: a bit stream can consist of interleaved sequences of diverse information packets. Some of these will generally be representations of some derived form of the audio, and the rest will be descriptive information about the signal, referred to as side-chain information.

Lossless Audio Coding

The goal of lossless coding is to convey digital information in a more compact form while reliably recovering the original data in the decoder with bit-for-bit accuracy. Obviously, no psychoacoustic or perceptual considerations need enter into the picture.

The key to the ability to losslessly compress digital data is redundancy. Any time there is a pattern or common element to a block of data, it may be possible to express it in a more compact form by conveying the essence of the pattern along with the deviations from the pattern, such that the original data may be reconstructed. Of course, in order for this to happen, the bit-stream syntax must support a compact means of specifying the necessary range of patterns to the decoder.

The number of potentially useful patterns is actually fairly large, and considerable ingenuity has gone into the development of lossless coding algorithms which can recognize and concisely encapsulate the details of redundant information patterns. Some of the more common data structures that can be losslessly compressed are listed below.

1. A repeating sample value: rather than send the same value multiple times, the sample can be sent just once, along with a replication factor. This technique is sometimes referred to as *run-length coding*.
2. Similar values: a variant of run-length coding, this type of pattern can be compressed by sending the average value, followed by concise representations of the difference between each sample value and the average. Alternatively, differential coding may be used, in which the difference between successive samples is coded.
3. Repeating sinusoids: for an ongoing sinusoidal signal, the PCM samples can be replaced with parameters indicating the amplitude, frequency, phase, and length of the sample sequence. In general form, this is sometimes referred to as *linear prediction*. If there are small deviations, they may be sent as concise difference values.
4. Small values: a sequence of samples, each with one or more leading zeros, may be sent with the zeros omitted, along with a factor indicating how many leading zeros were suppressed. (Note: negative values expressed in two's-complement form will have leading ones, but the same factor should otherwise apply.) This is known as *block floating point* representation.
5. Commonly used values: sample values which appear more often than the average – like the letters 'e'

Amazon EX1014 Page 780

and 's' in English text – may be assigned special, short transmission symbols to reduce the average data rate. Representative techniques which employ this include arithmetic coding and Huffman coding.
6. Interchannel difference coding: if two or more channels are used and are sufficiently similar, data-rate savings may be attained by sending sum and difference information, or a variant thereof. The expectation is that the difference values will be small, and can be coded in fewer bits using technique (4), above. In the case of stereo content, this approach is sometimes referred to as mid/side (M/S) coding.

In addition to the techniques described above (and others), a lossless coder may either directly code time-domain PCM samples, or may instead use running transforms and operate in the frequency domain. In the latter case, a specialized bit-exact transform may be used to guarantee that the final output is an exact clone of the input.

A fully realized lossless coder is likely to use a combination of techniques like those described above, requiring a flexible and perhaps elaborate bit stream syntax. The process of selecting which coding options will be brought to bear on a given block of samples is often not readily apparent from inspection or simple calculation, so the encoder may be required to try an exhaustive search of all possibilities for each block in order to select the best option. Needless to say, this can make for a rather slow-running encoder, but usually does not affect the speed of the decoder.

As may be evident, the effectiveness of a lossless coder is very much signal dependent: some signals can be compressed considerably more than others, and some, such as full-level white noise, basically cannot be losslessly compressed at all. Typical compression ratios run in the range of 2:1 to 3:1. Because lossless compression is inherently a variable bitrate (VBR) coding technology, it is not well suited to real-time streaming applications, such as broadcast or fixed-speed digital tape.

One weakness of lossless coding, and indeed of lossy coding as well, is that an error of even a single bit (in, say, one of the control codes specifying which method to use to unpack a block of data) can cause the loss of significant audio data. Some care must be used in the design of the bitstream to ensure that re-synchronization of the decoder in the face of data errors is reliably possible within an acceptably short period. More generally, the complexity and low error tolerance of almost any data-reduction algorithm raises the regrettable possibility that digital records, audio and otherwise, may be lost to future generations unless care is taken to avoid data errors, and to ensure that decoding algorithms are carefully documented and preserved in full detail. This is especially critical in the case of proprietary algorithms owned and maintained by commercial entities, whose long-term future operation may not be assured.

Perceptual/Lossy Audio Coding

While one might wish that lossless coding be used exclusively for digital audio compression, its limited performance and variable bitrate render it inappropriate for a large number of coding applications. For more-aggressive compression and acceptable performance within the constraint of constant bit rate, one must resort to the use of perceptual audio coding.

The goal of perceptual audio coding is to preserve and convey all of the perceptual aspects of audio content, without regard to preserving the signal literally, usually employing the fewest bits possible. Recall that the data rate for a standard audio CD was about 1.4 Mbit/s. As this is written, the lowest data rate in commercial use for conveying high-quality stereo audio is no higher than 48 Kbits/s, used by satellite radio and for Internet streaming, a data-compression ratio on the order of 30:1 or better.

How is this accomplished? For starters, the use of lossless coding is not retained and incorporated in overall lossy algorithms. Every packet of information generated and reduced to its ultimate compactness by any part of a perceptual coder is a candidate for further reduction in size if one or more lossless techniques can be brought to bear effectively.

Beyond that, the keys to high-efficiency perceptual coding are to:

1. suppress any audio components not contributing to the ultimate perception,
2. ensure that all the audible components are conveyed with no alteration in perceived quality, while using the minimum possible number of bits, and
3. employ, as much as possible, high-level abstractions to describe audio signals in purely perceptual terms, while retaining enough precision and detail to allow the decoder to accurately reconstitute a signal indistinguishable from the original encoder input signal.

One fundamental building block that has almost always been part of the foundation of perceptual coders is a digital filter bank. This, of course, mirrors the functionality of the basilar membrane; and the frequency-domain out-

Amazon EX1014 Page 781

put signals are most commonly used for both the analysis sections of the coder and the signal path. Indeed, the fundamental data conveyed by most perceptual coders is not PCM time-domain data, but quantized frequency-domain information from which the decoder eventually produces PCM output by way of a synthesis filter bank. The most commonly used digital filter bank is a variant of the FFT called the time-domain alias cancellation (TDAC) transform, developed by Princen and Bradley in 1986 [18.69, 70]. This transform has the highly desirable property of being critically sampled, meaning that it produces exactly the same number of output frequency-domain samples as there are input PCM samples. There is also a companion inverse TDAC transform for producing output PCM from a set of decoded frequency-domain input samples.

In operation, PCM samples input to the encoder are divided into regular blocks of some predetermined length, and each block is transformed to the frequency domain, conveyed to the decoder, reconstituted as a block of PCM samples, and the successive blocks strung together, often in an overlapping manner, to recover the final outputs. This process is sometimes described as block-oriented processing using running transforms. The practice of coding an entire block of samples as a single entity facilitates data rates of less than one bit per original audio sample. One downside of block processing is that events which occupy a small fraction of a block, like a sharp transient, may become smeared across a range of frequencies, consuming a large number of bits. A number of techniques have evolved to deal with this situation, most commonly the use of block switching, wherein such short events are detected (by, for example, a transient detector routine), and the block size is temporarily shortened to more closely isolate the transient within the shortened block.

The filter-bank output is of direct use in pursuit of the first two perceptual coding techniques listed above, principally by deriving the magnitude of the frequency-domain signals as a function of frequency to obtain a discrete approximation to the power spectral density (PSD) of the signal block. This in turn is processed by a routine implementing a perceptual model of human hearing to derive a masking curve, which specifies the threshold of hearing as a function of frequency for that signal block. Any spectral component falling below the masking curve will be inaudible, so need not be coded, as per the first listed coding technique. The remaining spectral components must be preserved if audible alteration is to be avoided, but the quantization precision is only that which is required to render the level of the quantization noise below the masking curve. Instead of the 120 dB/20 bit range one might require for PCM audio, the instantaneous masking range is more often on the order of 20–30 dB, about 4–5 bits per sample, assuming 6 dB per bit SNR. Thus, between suppression of inaudible components and dynamic quantization of audible components, the data rate can be expected to be reduced to something less than 4 bits per sample. Of course, an effective bit-stream syntax protocol must be devised to signal these conditions efficiently to the decoder, and considerable ingenuity has been brought to bear on that issue to ensure the requirement is met.

Notable lossy audio coders based on these principles include AC-3, MP3, DTS, WMA, Ogg, and AAC. These have been instrumental enablers of such technologies as portable music players, DVD's, digital soundtracks on 35 mm film, and satellite radio.

It should be noted that perceptual coding is much more compatible with constant-bit-rate operation than is lossless coding, for as the complexity of a signal increases, which might otherwise increase the required data rate, the masking afforded by that signal also increases, reducing the average quantization accuracy required, thereby holding the required data rate to a more nearly constant rate. In effect, the human auditory system can only absorb so much information per unit time, so as long as the coder accurately models that behavior, the required data rate should be largely constant. Of course, very simple signals, like silence, are not likely to require the same data rate as more complex signals, in which case a constant-bitrate coder may simply use far more than the minimum required data rate in order to maintain a constant transmission rate. Some coders can defer use of bits in the presence of simple signals, saving them in a *bit bucket* until needed by a more complex signal.

The third listed technique used in perceptual coding, the application of higher-level abstractions to describe the signal, is much more general and open-ended; and is still an area of active investigation. It can be something as simple as using a single PSD curve to approximate the actual PSD curves of several successive transform blocks, or using decoder-generated noise in place of actually transmitting noise-like signals. More abstract approaches may fall under the heading of *parametric coding*.

One notable example of parametric coding is *bandwidth extension*, in which the entire high-frequency end of the spectrum is suppressed by the encoder, and is reconstituted by the decoder from analysis of the harmonic

Amazon EX1014 Page 782

content of rest of the signal spectrum. This technique alone can reduce the data rate by 30% or more.

Another widely pursued approach to parametric coding is to segregate signal components into categories, such as sinusoidal, noise-like, or transient, then convey numerical descriptions of each to a synthesizing decoder. Although this is arguably far closer to a true perceptually grounded coder than a simple transform-based coder, the complexity of the descriptors needed to accurately portray the full range of possible signals has so far made it difficult to reduce the data rate required by more conventional coders dramatically.

Spatial Audio Coding

Although strictly speaking it may be a subset of perceptual audio coding, spatial audio coding (SAC) is a sufficiently independent approach to justify separate examination. SAC can be used to improve the transmission efficiency of multichannel content, typically either stereo or 5.1-channel surround material.

The approach involves coding the ensemble of channels as a single entity to achieve data savings that would be unavailable if each channel were to be coded individually. An element common of many SAC algorithms is to combine the source channels additively to a smaller number of channels, perhaps as few as a single channel, from which the decoded output channels are eventually reconstructed. The data rate is thereby reduced by the ratio of the number of input channels to the number of down-mixed channels, minus any side-chain information needed by the decoder to reconstruct the output channels.

Probably the simplest approach to SAC is to downmix the input channels to mono in the encoder and then simply distribute the common signal equally to the output channels. If this is done to the full-bandwidth signal, the result will simply be monaural sound reproduced from multiple speakers, which is not likely to be a very satisfactory multichannel experience, but if limited to a subset of the audio range, say just high frequencies, the ear will tend to associate the spread mono content of the reproduced channels with the direction of the lower frequency, discrete content, and the alteration will be less noticeable.

A slightly more elaborate approach is to send channel-by-channel scale factors, representing relative signal amplitude in each input channel, which are subsequently used by the decoder to distribute signal components to each output channel. Sometimes referred to as either intensity stereo or channel coupling, the technique can be surprisingly effective, especially at high frequencies when individual scale factors are sent on something like a critical band frequency basis.

More recent spatial coding systems have augmented the use of amplitude coupling with transient flags and information about interchannel phase and coherence, allowing extension to progressively lower frequencies, with commensurately greater amounts of data rate reduction.

Although in most cases the spatial information about individual channels is sent as side-chain information comprising part of the total transmitted bit stream, there has been some success in encoding the information in the signal characteristics of the actual transmitted audio, which can reduce the net data rate.

Spatial audio coding remains at this time an area of very active investigation.

18.6 Complete Audio Systems

So far, this discussion of audio and electroacoustics has focused on the task of capturing, transmitting, and reproducing one or more individual channels of audio as accurately as possible. However, the goal of sound reproduction is not strictly to convey some designated number of channels, but to faithfully reproduce a full acoustic event, be it live, synthesized, or some combination thereof, to an audience that, at any given venue, may range in number from one to many hundreds or more.

The use of individual channels is mandated by the available technology. Microphones and loudspeakers are basically point-in-space devices, and most audio media and processors are designed to handle individual channels. But sound is, by nature, a three-dimensional entity, not counting time and frequency, and a few isolated channels can at best represent a sparse spatial sampling of a full acoustic sound field.

So the question is raised as to just how effective are complete audio systems, from recording to reproduction, at least in perceptual, psychoacoustic terms.

Any attempt to perform such an evaluation is immediately faced with the conundrum as to just what should be the ultimate goal of sound reproduction. Should we be trying to transport the listener to the performance site ("we are there"), or is the goal to transport the performers to the listener's living room ("they are here").

Amazon EX1014 Page 783

The answer perhaps depends on the nature of the source material, as well as listener preference. For situational acoustics, from walking down a busy street to riding in a submarine, it would seem logical to go with "we are there". This is most commonly the case for, say, cinema soundtracks, making a good case for surround-sound systems in commercial and home theaters. A similar case can be made for symphonic music, since the concert hall is such an integral part of the experience. However, a small ensemble or a solo artist might be more compelling if holographically projected into the listening room. Given unlimited wishes, one might hope to have both, with the ability choose which option according to one's predilections. In actual practice, we might be well served if either one is possible. But in any case, it is perhaps worth keeping in mind in evaluating the effectiveness of complete audio systems.

18.6.1 Monaural

For the first half century of recorded sound, monaural was all there was. In that time, considerable expertise was developed in microphone techniques for capturing sound with a viable balance of direct and ambient content. In this era of multichannel audio, it is easy to dismiss mono for failing to support any formal spatial audio cues. Still, there are times when a single channel should, at least in theory, be adequate for the task, such as a solo performer or other single, localized source, in a dry (relatively echo-free) ambient environment, or "here". Even in that case, however, the reproducing loudspeaker is unlikely to have the same spatial radiation pattern as the original source, a problem common to many sound formats to one degree or another, so that even with good frequency response and low distortion, the sound may be recognizable as emanating from a loudspeaker, rather than being a clone of the original performance. To its credit, mono does not exhibit the double-arrival problem of phantom images common to stereo and other multichannel formats (below), arguably resulting in a purer, less time-smeared sound. It is also gratifyingly free of sweet-spot dependence: one can sit pretty much anywhere in a listening room without worrying about impairment of imaging. And historically at least, mono was a big improvement over nothing at all.

18.6.2 Stereo

Two-channel stereophonic sound has been a commercial reality for nearly half a century, and shows little sign of being rendered obsolete anytime soon. The intuitive attraction of stereo is self-evident: we have two ears, and stereo provides two channels. On the surface, this is an apparently compelling argument. The result should perfect sound reproduction, yes? Well, no, although stereo did pretty much render mono obsolete for most media within a few years of its commercial deployment.

Stereo provides the smallest possible increment in spatial fidelity over mono: a horizontal line spread before the listener. But this is likely the most important single improvement one could hope to make, as the human ear is most sensitive to front-center horizontal position. In addition, performing ensembles are most often arrayed in such a manner. The arrangement also allows rendering, where appropriate, of frontal ambiance in a manner more natural than can be achieved with a single audio channel.

An important element of stereo reproduction is the creation of virtual images. These are sonic images that appear suspended in space between the two speakers, and they undoubtedly add considerably to the viability of the format. Virtual images depend for their existence on the primary horizontal localization cues: interaural time differences and interaural amplitude differences. The former cue is especially critical to the creation of virtual images, as the total range of the cue is only about $\pm 700\,\mu s$, corresponding to a horizontal displacement of the listener of only about a foot. This rather undermines the viability of stereo virtual imaging somewhat, and renders the format decidedly sweet-spot dependent. The situation is partially compensated by the flexibility of the human auditory system, and its apparent ability to use dynamically whatever cues appear to be providing reliable localization information. A listener a little off the center line may still experience some degree of virtual imaging by virtue of the interaural amplitude cue, which may not degrade as rapidly as the interaural time delay cue. Indeed, in the course of mixing a multichannel source to a stereo track, a recording engineer will typically position a virtual source between the speakers using simple amplitude panning, with zero interaural time delay, and yet the ear generally accepts this.

But sensitivity to listener position is not the only issue relating to virtual images. There is also that double-arrival problem. The direct sound from almost any real sound source will generally arrive at each ear once, and that is it. But a virtual image from a stereo speaker array will result in two direct arrivals at each ear, one from each speaker. Assuming a centered listener, the left ear will hear from the left speaker at about the same time that the right ear hears the first arrival from the right speaker. Then the sound from each speaker will

Amazon EX1014 Page 784

cross the head, in opposite directions, and the left ear will hear the right speaker arrival, and vice versa, a process sometimes referred to as loudspeaker crosstalk. The time delay between arrivals is typically on the order of 250 μs, resulting in a small, subtle, but audible notch in the resulting response around 2 kHz. Some temporal smearing of sharp transients may also be evident.

A final difficulty with virtual imaging of stereo systems is the relative inability to position virtual images outside the range of the speakers. This is another consequence of loudspeaker crosstalk, limiting the maximum interaural time difference to that corresponding to the position of the speaker, which again is about 250 μs, and similarly limiting the maximum interaural amplitude difference. One way around this would be to put the speakers in different rooms and sit between them, one ear in each room, with a custom-fitted baffle doorway to block any crosstalk, which is a little impractical. Another approach is to employ a crosstalk cancellation (virtual-speaker) system, to cancel the second arrival at each ear, but such systems are themselves very sweet-spot dependent, and even then usually limited to creating images on the horizontal plane. The most direct way to provide sound from additional directions is to position additional speakers around the listener, including above and possibly below, assuming additional source channels are available to feed them, but that is no longer a stereo sound system.

18.6.3 Binaural

Like stereophonic sound, binaural makes use of just two channels, but it uses headphones instead of loudspeakers to avoid the crosstalk problem. On paper, this system seems like it should be even more perfect than stereo. However, the effect is often to place the sound images within the head, with little or no front/back differentiation.

Part of the problem appears to be that the presentation does not get altered when the listener's head moves, so that the image follows the head movement, which is rather unnatural. There has been some work exploring the use of tracking headphones, which alter the signal in response to head motion, but these have not yet been completely successful, and require more than two channels to be available.

Another issue with binaural reproduction seems to be the lack of pinna cues, those alterations in timing and spectra that are imparted by the pinnae to an arriving sound as a function of direction. Efforts to impart pinna cues electronically have had some success, but they tend to be as individual as fingerprints, making it difficult to have a single binaural track which works for everyone. There has also been some success in producing a universal binaural track by imparting a direction-dependent room reverberation characteristic to the signals, although some purists object to the addition, and the effect is still somewhat listener dependent.

One experimental system that mitigates some of the shortcomings of two-channel binaural is four-channel binaural. This employs two dummy heads or equivalently arranged directional microphones, with one head/microphone-pair in front of the other, preferably with an isolating baffle separating them. The resulting four-channel recording is played over four speakers, preferably placed close to the listener in fixed positions around the listener's head, corresponding to the microphone positions, possibly with a left–right isolating panel or crosstalk-canceling circuits to minimize crosstalk. Because the front–back pair each preserve ITD and IAD, the left–right space is properly preserved, while front–back differentiation and externalization are provided by using separate front and back external drivers. Dynamic IAD and ITD cues with head motion are also at least crudely preserved by the use of front and back channels. The resulting presentation can exhibit good spatial fidelity.

Although it may never be suitable for general audience presentation, the binaural format has compelling qualities for the single listener, and remains an area of active investigation.

18.6.4 Ambisonics

This system, devised by Michael Gerzon, approximates the sound field in the vicinity of a listener using an array of precisely positioned loudspeakers, It is sweet-spot dependent, so not amenable to audience presentation, but can provide compellingly good spatial reproduction.

18.6.5 5.1-Channel Surround

The 5.1-channel format is a well-thought-out response to the shortcomings of stereo. The format employs three front speakers, two surround speakers placed to the side and/or back of the listening room, and the 0.1 low-frequency channel that only occupies a small fraction of the total bandwidth.

The addition of the center-front speaker anchors the front left–right spread, significantly mitigating the problems stereo has with virtual images, even with audience presentation. The surround channels provide sound more

to the side than can be achieved with the front speakers (or stereo), exercising the full range of the ITD/IAD sensitivity of the ear, and providing the option of a far more immersive experience than can be attained with stereo.

The 5.1 format with high-quality source material provides the most accurate overall audio reproduction of any widely used format to date, within either the "here" or "there" reproduction paradigms.

18.7 Appraisal and Speculation

We have been witness, from the crude but noble origins of sound reproduction, to an astonishing march of progress that has, by degrees, rolled down the problems and imperfections of sound reproduction, reaching a notably high level of performance, when considered in basic perceptual terms.

A cynic might note that, fifty years ago, children played little yellow plastic 78 RPM 17.78 cm discs on phonographs that had 12.7 cm loudspeakers, and now, after fifty years of dramatic, hard-won progress, many people play little 12.7 cm plastic (compact) discs on boom boxes with 12.7 cm loudspeakers. Yes, the fidelity is better, as is the playing time, but have we really come that far?

The apparent answer from all the above is that a high-quality system of modern vintage can now satisfy a major percentage of the electroacoustic and spatial requirements implied by the capabilities of the human auditory system.

This does not mean that audio engineers call all just pack up and go home just yet. Indeed, the quantities of audio research papers and patents published each year show no sign of abating, giving rise to the speculation that further meaningful improvements may yet be attained.

One can certainly wish for a few more channels, to fill in the remaining gaps and perhaps even provide explicit coverage above and below the listener. The issue in that regard at this point is more of practicality than of theory. There have been some experiments with the addition of height or ceiling (*voice of God*) channels, and the specifications for digital cinema systems allow for the use of over a dozen channels.

One can also hope for enhanced ability to project sound into a room, rather than having it just arrive from the walls. That is another area of active investigation. One approach, referred to as wave field synthesis [18.71] has been to approximate a complete sound field by reproducing the sound-pressure profile around a periphery, and relying on the Huygens principle [18.72] to recreate the complete interior sound field. This requires large number of loudspeakers (192 in one experimental setup) to avoid spatial aliasing effects.

The other well-known approach to generating interior sounds is to use an ultrasonic audio projector, which projects acoustic signals above the audio band which heterodyne to produce audible sound which appears to come from a remote location [18.73].

Another problem that has only partly been solved is that of sweet-spot sensitivity. This has long been an issue with cinema surround-sound systems, but is no less challenging for automotive surround systems, where there are often no centrally located seats and speaker locations may be less than optimal. The issues go beyond balanced sound to trying to provide a coherent, immersive experience at all listening positions.

Indeed, each new venue for audio seems to come with its own idiosyncratic problems, leading to active investigation of how to make compelling presentations for portable music players, cell phones, and interactive media such as game consoles.

It is also expected that there will be continuing refinement of binaural techniques for personal listening.

Ultimately there may come a wholesale revolution in audio techniques, perhaps with laser transducers replacing loudspeakers and microphones, along with holographic recording of entire sound fields. For personal listening, your music player might interface directly to your brain, bypassing the ear altogether, allowing otherwise deaf people to hear, and getting rid of that troublesome 3 kHz peak in the ear canal once and for all.

It should make for some interesting vibes.

References

18.1 N. Aujoulat: *The Cave of Lascaux* (Ministry of Culture and Communication, Paris August 2005), http://www.culture.gouv.fr/culture/arcnat/lascaux/en/

Amazon EX1014 Page 786

18.2 J. Sterne: *The Audible Past* (Duke Univ. Press, Durham 2003)
18.3 Thomas Edison (Wikipedia Aug. 2005) http://en.wikipedia.org/wiki/Thomas_Edison
18.4 J.H. Lienhard: *The Telegraph, Engines of Our Ingenuity* (Univ. Houston, Houston 2004), http://www.uh.edu/engines/epi15.htm
18.5 P. Ament: The Electric Battery (The Great Idea Finder, Mar. 2005) http://www.ideafinder.com/history/inventions/story066.htm
18.6 Phonograph (Wikipedia, Aug. 2005) http://en.wikipedia.org/wiki/Phonograph#History
18.7 D. Marty: The History of the Phonautograph (Jan. 2004) http://www.phonautograph.com/
18.8 Emile Berliner and the Birth of the Recording Industry, Library of Congress, Motion Picture, Broadcasting and Recorded Sound Division (April, 2002) http://memory.loc.gov/ammem/berlhtml/berlhome.html
18.9 J. Bruck, A. Grundy, I. Joel: *An Audio Timeline* (Audio Engineering Society, New York 1997), http://www.aes.org/aeshc/docs/audio.history.timeline.html
18.10 M.F. Davis: History of spatial coding, J. Audio Eng. Soc. **51**(6), 554–569 (2003)
18.11 NIH: *Hearing Loss and Older Adults, NIH Publication No. 01-4913, National Institute on Deafness and Other Communications Disorders* (National Inst. Health, Bethesda 2002), http://www.nidcd.nih.gov/health/hearing/older.asp
18.12 T. Weber: *Equal Loudness* (WeberVST, Kokomo 2005), http://www.webervst.com/fm.htm
18.13 E. Terhardt: *Pitch Perception* (Technical Univ. Munich, Munich 2000), http://www.mmk.e-technik.tu-muenchen.de/persons/ter/top/pitch.html
18.14 B. Truax: *Sound pressure level [SPL], Handbook for Acoustic Ecology, Simon Fraser University, Vancouver* (Cambridge Street Publ., Burnaby 1999), http://www2.sfu.ca/sonic-studio/handbook/Sound_Pressure_Level.html
18.15 B. Truax: *Phon, Handbook for Acoustic Ecology, Simon Fraser University, Vancouver* (Cambridge Street Publ., Burnaby 1999), http://www2.sfu.ca/sonic-studio/handbook/Phon.html
18.16 B.G. Crockett, A. Seefeldt, M. Smithers: *A new loudness measurement algorithm, AES 117th Convention Preprint 6236* (Audio Engineering Society, New York 2004)
18.17 L.H. Carney: *Studies of Information Coding in the Auditory Nerve* (Syracuse Univ., Syracuse 2005)
18.18 E.D. Haritaoglu: *Temporal masking* (University of Maryland, College Park 1997), http://www.umiacs.umd.edu/ desin/Speech1/node11.html
18.19 K.M. Steele: *Binaural processing* (Appalachian State Univ., Boone April 2003), http://www.acs.appstate.edu/kms/classes/psy3203/SoundLocalize/intensity.jpg
18.20 J. Blauert: *Spatial Hearing* (MIT Press, Cambridge 1983)
18.21 R. Kline: Harold Black and the negative-feedback amplifier, IEEE Control Syst. Mag. **13**(4), 82–85 (1993), DOI: 10.1109/37.229565
18.22 N. Moffatt: *Wow and Flutter* (Vinyl Record Collectors, United Kingdom 2004), http://www.vinylrecordscollector.co.uk/wow.html
18.23 B. Paquette: *History of the microphone* (Belgium 2004), http://users.pandora.be/oldmicrophones/microphone_history.htm
18.24 J. Strong: *Understanding microphone types, Pro Tools for Dummies* (Wiley, New York 2005), http://www.dummies.com/WileyCDA/DummiesArticle/id-2509.html
18.25 P.V. Murphy: *Carbon Microphone* (Concord University, Athens 2003), http://students.concord.edu/murphypv/images/35107.gif
18.26 D.A. Bohn: *Ribbon Microphones* (Rane Corp., Mukilteo 2005), http://www.rane.com/par-r.html
18.27 H. Robjohns: *A brief history of microphones, Rycote Microphone Data Book* (Rycote Microphone Windshields, Stroud 2001), http://www.microphone-data.com/pdfs/History.pdf (Rycote Microphone Windshields, Stroud 2003)
18.28 G. M. Sessler, J. E. West: Electroacoustic Transducer Electret Microphone, US Patent Number 3118022 (1964) http://www.invent.org/hall_of_fame/132.html (National Inventors Hall of Fame, 2002)
18.29 P.V. Murphy: *Piezoelectric Microphone* (Concord University, Athens 2003), http://students.concord.edu/murphypv/images/35108.gif
18.30 D. Stewart: *Understanding microphone self noise* (ProSoundWeb, Niles 2004), http://www.prosoundweb.com/studio/sw/micnoise.php
18.31 A.G.H. van der Donk, P.R. Scheeper, P. Bergveld: Amplitude-modulated electro-mechanical feedback system for silicon condenser microphones, J. Micromech. Microeng. **2**, 211–214 (1992), http://www.iop.org/EJ/abstract/0960-1317/2/3/024
18.32 A. Oja: *Capacitive Sensors and their readout electronics, Helsinki Univ. Technology, Physics Laboratory* (Helsinki Univ. Technology, Espoo 2005), http://www.fyslab.hut.fi/kurssit/Tfy3.480/Spring2004/Jan28_Oja/Microsensing_2004_Oja2.pdf
18.33 J.N. Caron: *Gas-coupled laser acoustic detection* (Quarktet, Lanham 2005), http://www.quarktet.com/GCLAD.html
18.34 H.M. Ladak: *Anatomy of the middle ear, Finite-Element Modelling Of Middle- Ear Prostheses in Cat, Master of Engineering Thesis* (McGill University, Montreal 1993), Chap. 2 http://audilab.bmed.mcgill.ca/ funnell/AudiLab/ladak_master/chapter2.htm

Amazon EX1014 Page 787

18.35 Microphone Types, The Way of the Classic CD (Musikproduktion Dabringhaus und Grimm, Detmold 2005), http://www.mdg.de/mikroe.htm

18.36 G. Carol: *Tonearm/Cartridge Capability* (Galen Carol Audio, San Antonio 2005), http://www.gcaudio.com/resources/howtos/tonearmcartridge.html

18.37 J.G. Holt: *Down With Dynagroove, Stereophile, December 1964* (Primedia Magazines, New York 1964), http://www.stereophile.com/asweseeit/95/

18.38 D. Krotz: *From Top Quarks to the Blues, Berkeley Lab physicists develop way to digitally restore and preserve audio recordings, Berkeley Lab. Res. News 4/16/2004* (Lawrence Berkeley National Lab., Berkeley 2004), http://www.lbl.gov/Science-Articles/Archive/Phys-quarks-to-blues.html

18.39 S.E. Schoenherr: *Loudspeaker history, Recording Technology History* (Univ. San Diego, San Diego July 2005), http://history.sandiego.edu/gen/recording/loudspeaker.html

18.40 C.W. Rice, E.W. Kellogg: Notes on the development of a new type of hornless loudspeaker, Trans. Am. Inst. El. Eng. **44**, 461–475 (1925)

18.41 AudioVideo 101: *Speakers Enclosures* (AudioVideo 101, Frisco 2005), http://www.audiovideo101.com/learn/knowit/speakers/speakers16.asp

18.42 Eminence: *Understanding loudspeaker data* (Eminence Loudspeaker Corp., Eminence 2005), http://editweb.iglou.com/eminence/eminence/pages/params02/params.htm

18.43 M. Sleiman: *Speaker Engineering Analysis* (Northern Illinois University, DeKalb 2005), www.students.niu.edu/ z109728/UEET101third%20draft.doc

18.44 Allison Acoustics: *Company History* (Allison Acoustics LLC, Danville 2005), http://www.allisonacoustics.com/history.html

18.45 D. Bohn: A fourth-order state-variable filter for Linkwitz-Riley active crossover designs, J. Audio Eng. Soc. Abstr. **31**, 960 (1983), preprint 2011

18.46 W. Norris: *Overview of hypersonic technology* (American Technology Corp., San Diego 2005), http://www.atcsd.com/pdf/HSSdatasheet.pdf

18.47 T. Sporer: *Conveying sonic spaces* (Fraunhofer Elektronische Medientechnologie, Ilmenau 2005), http://www.emt.iis.fhg.de/projects/carrouso/

18.48 B. DePalma: Analog Audio Power Amplifier Design (January, 1997) http://depalma.pair.com/Analog/analog.html

18.49 E. Augenbraun: *The Transistor, Public Broadcasting System* (ScienCentral, The American Institute of Physics, New York 1999), http://www.pbs.org/transistor (July 2005)

18.50 S. Portz: *Who Invented The Transistor* (PhysLink, Long Beach 2005), http://www.physlink.com/Education/AskExperts/ae414.cfm

18.51 THAT Corp.: *A brief history of VCA development* (THAT Corp., Milford 2005), http://www.thatcorp.com/vcahist.html

18.52 M. Hart: *History of Sound Motion Pictures* (American WideScreen Museum, 2000), http://www.widescreenmuseum.com/sound/sound03.htm

18.53 L.P. Lessing: *Edwin H. Armstrong, Dictionary of American Biography, Supplement V* (Scribner, New York 1977) pp. 21–23, http://users.erols.com/oldradio/ (2005)

18.54 A. Antoniou: *Digital Filters: Analysis, Design, and Applications*, 2nd edn. (McGraw-Hill, New York 1993)

18.55 A. Antoniou: *On the Origins of Digital Signal Processing* (Univ. Victoria, Victoria 2004), http://www.ece.uvic.ca/ andreas/OriginsDSP.pdf

18.56 R.W.T. Preater, R. Swain: Fourier transform fringe analysis of electronic speckle pattern interferometry fringes from hinge-speed rotating components, Opt. Eng. **33**, 1271–1279 (1994)

18.57 E.O. Bringham: *The Fast Fourier Transform* (Prentice-Hall, Upper Saddle River 1974)

18.58 K.C. Pohlmann: *The Compact Disc Handbook*, Comp. Music Digital Audio Ser., Vol. 5, 2nd edn. (A-R Editions, Middleton 1992)

18.59 J. Despain: *History of CD Technology* (OneOff Media, Salt Lake City 1999), http://www.oneoffcd.com/info/historycd.cfm?CFID=435613&CFTOKEN=10955739

18.60 E.T. Whittaker: On the functions which are represented by the expansions of interpolation-theory, Proc. R. Soc. Edinburgh. **35**, 181–194 (1915)

18.61 C.E. Shannon: A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379–423 (1948)

18.62 C.E. Shannon: A mathematical theory of communication, Bell Syst. Tech. J. **27**, 623–656 (1948)

18.63 E. Meijering: Chronology of interpolation, Proc. IEEE **90**(3), 319–342 (2002), http://imagescience.bigr.nl/meijering/research/chronology/

18.64 M.T. Heideman, D.H. Johnson, C.S. Burrus: Gauss and the history of the fast Fourier transform, IEEE Acoust. Speech Signal Process. Mag. **1**, 14–21 (1984),

18.65 M.T. Heideman, D.H. Johnson, C.S. Burrus: Gauss and the history of the fast Fourier transform, Arch. Hist. Exact Sci. **34**, 265–277 (1985)

18.66 C.S. Burrus: *Notes on the FFT* (Rice University, Houston 1997), http://www.fftw.org/burrus-notes.html

18.67 P. Duhamel, M. Vetterli: Fast Fourier transforms: A tutorial review and a state of the art, Signal Process. **19**, 259–299 (1990)

18.68 Fast Fourier Transform (Wikipedia, August 2005) http://en.wikipedia.org/wiki/Fast_Fourier_transform

18.69 M.F. Davis: *The AC-3 multichannel coder, Audio Engineering Society. 95th Convention 1993* (Audio Engineering Society, New York 1993), Preprint Number 3774

18.70 M.F. Davis: *The AC-3 multichannel coder* (Digital Audio Development, Newark 1993), http://www.dadev.com/ac3faq.asp

Amazon EX1014 Page 788

18.71 T. Sporer: *Creating, assessing and rendering in real time of high quality audio- visual environments in MPEG-4 context, Information Society Technologies* (CORDIS, Luxemburg 2002), http://www.cordis.lu/ist/ka3/iaf/projects/carrouso.htm

18.72 G. Theile, H. Wittek: Wave field synthesis: A promising spatial audio rendering concept, Acoust. Sci. Tech. **25**, 393–399 (2004), DOI: 10.1250/ast.25.393

18.73 E.I. Schwartz: *The sound war, Technol. Rev. May 2004* (MIT, Cambridge 2004), http://www.woodynorris.com/Articles/TechnologyReview.htm

Amazon EX1014 Page 789

Part F Biological and Medical Acoustics

19 Animal Bioacoustics
Neville H. Fletcher, Canberra, Australia

20 Cetacean Acoustics
Whitlow W. L. Au, Kailua, USA
Marc O. Lammers, Kailua, USA

21 Medical Acoustics
Kirk W. Beach, Seattle, USA
Barbrina Dunmire, Seattle, USA

Amazon EX1014 Page 790

19. Animal Bioacoustics

Animals rely upon their acoustic and vibrational senses and abilities to detect the presence of both predators and prey and to communicate with members of the same species. This chapter surveys the physical bases of these abilities and their evolutionary optimization in insects, birds, and other land animals, and in a variety of aquatic animals other than cetaceans, which are treated in Chap. 20. While there are many individual variations, and some animals devote an immense fraction of their time and energy to acoustic communication, there are also many common features in their sound production and in the detection of sounds and vibrations. Excellent treatments of these matters from a biological viewpoint are given in several notable books [19.1, 2] and collections of papers [19.3–8], together with other more specialized books to be mentioned in the following sections, but treatments from an acoustical viewpoint [19.9] are rare. The main difference between these two approaches is that biological books tend to concentrate on anatomical and physiological details and on behavioral outcomes, while acoustical books use simplified anatomical models and quantitative analysis to model whole-system behavior. This latter is the approach to be adopted here.

19.1 **Optimized Communication** 785
19.2 **Hearing and Sound Production** 787
19.3 **Vibrational Communication** 788
19.4 **Insects** 788
19.5 **Land Vertebrates** 790
19.6 **Birds** 795
19.7 **Bats** 796
19.8 **Aquatic Animals** 797
19.9 **Generalities** 799
19.10 **Quantitative System Analysis** 799
References 802

19.1 Optimized Communication

Since animals use their acoustic and vibrational senses both to monitor their environment and to communicate with other animals of the same species, we should expect that natural selection has optimized these sensing and sound production abilities. One particular obvious optimization is to maximize the range over which they can communicate with others of the same species. Simple observation shows that small animals generally use high frequencies for communication while large animals use low frequencies – what determines the best choice? The belief that there is likely to be some sort of universal scaling law involved goes back to the classic work of *D'Arcy Thompson* [19.10], while a modern overview of physical scaling laws (though not including auditory communication) is given by *West* and *Brown* [19.11].

The simplest assumption is that the frequency is determined simply by the physical properties of the sound-producing mechanism. For a category of animals differing only in size, the vibration frequency of the sound-producing organ depends upon the linear dimensions of the vibrating structure, which are all proportional to the linear size L of the animal, and upon the density ρ and elastic modulus E of the material from which it is made. We can thus write that the song frequency $f = A\rho^x E^y L^z$, where A, x, y, and z are constants. Since the dimensions of each side of the equation must agree, we must have $x = -1$, $y = 1$ and $z = -1$, which leads to the conclusion that song frequency should be inversely proportional to the linear size of the animal or, equivalently, that $f \propto M^{-1/3}$, where M is the mass of the animal. This simple result agrees quite well with

Amazon EX1014 Page 791

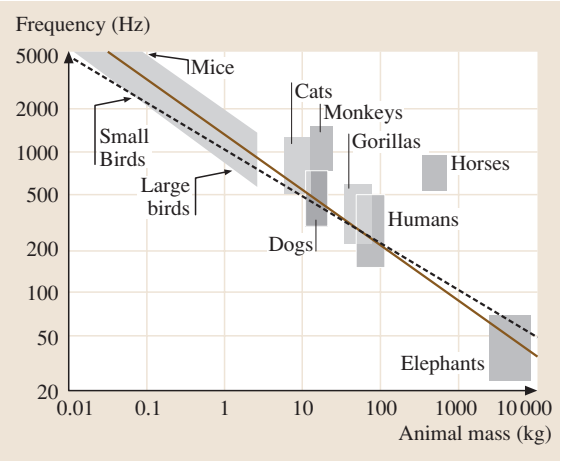


Fig. 19.1 The frequency ranges of the emphasized frequencies of vocalization in a large range of land-dwelling animals, plotted as a function of the mass of the animal. The *dashed line* shows the regression $f \propto M^{-0.33}$ while the *solid line* is the regression $f \propto M^{-0.4}$, as discussed in the text [19.12]

observation, as shown by the dashed line in Fig. 19.1, for a wide variety of animals, but a more detailed analysis is desirable.

Over reasonable distances in the air, sound spreads nearly hemispherically, so that its intensity decays like $1/R^2$, where R is the distance from the source. But sound is also attenuated by atmospheric absorption, with the attenuation coefficient varying as f^2. An optimally evolved animal should have maximized its communication distance under the influence of these two effects. An analysis [19.12] gives the result that $f \propto M^{-0.4}$ which, as shown in Fig. 19.1, fits the observations even better than the simpler result. There are, however, many outliers even among the animals considered, due to different anatomies and habitats. Insects, which must produce their sound in an entirely different way as is discussed later, are not shown on this graph, but there is a similar but not identical relative size relation for them too.

The total sound power produced by an animal is also a function of its size, typically scaling about as $M^{0.53}$ for air-breathing animals of a given category. When the variation of song frequency is included, this leads [19.12] to a conspecific communication distance proportional about to $M^{0.6}$. Again, while this trend agrees with general observations, there are many very notable outliers and great differences between different types of animals. Thus, while mammals comparable in size with humans typically produce sound power in the range 0.1–10 mW, and large birds may produce comparable power, some small insects such as cicadas of mass not much more than 1 g can also produce almost continuous calls with a power of 1 mW, as will be discussed in Sect. 19.4. At intermediate sizes, however, many animals, particularly reptiles, are almost mute.

Elephants represent an interesting extreme in the animal world because of their very large mass – as much as 10 tonnes. Their calls, which have a fundamental in the range 14–35 Hz, can have an acoustic power as large as 5 W, leading to a detection distance as large as 5 km, or even up to 10 km after sunset on a clear night [19.13] when very low frequency propagation is aided by atmospheric inversion layers (Chap. 4). Vegetation too, of course, can have a significant effect upon transmission distance. These elephant calls are often misleadingly referred to as "infrasonic" in the biological literature, despite the fact that only the fundamental and perhaps the second harmonic satisfy this criterion, and the calls have many harmonics well within the hearing range of humans. Indeed, even other elephants depend upon these upper harmonics to convey information, and usually cannot recognize the calls of members of the same family group above a distance of about 1.5 km because of the attenuation of harmonics above about 100 Hz through atmospheric absorption [19.14].

When it comes to sound detection, rather similar scaling principles operate. As will be discussed briefly in Sect. 19.2, the basic neuro-physiological mechanisms for the conversion of vibrations to neural impulses are remarkably similar in different animal classes, so that it is to be expected that auditory sensitivity should vary roughly as the area of the hearing receptor, and thus about as $M^{2/3}$, and this feature is built into the analysis referred to above. While a few animals have narrow-band hearing adapted to detecting particular predators – for example caterpillars detecting wing beats from wasps – or for conspecific communication, as in some insects, most higher animals require a wide frequency range so as to detect larger predators, which generally produce sounds of lower frequency, and smaller prey, which may produce higher frequencies. The auditory range for most higher animals therefore extends with reasonable sensitivity over a frequency range of about a factor 300, with the central frequency being higher for smaller animals. In the case of mammals comparable in size to humans, the range for mature adults is typically about 50 Hz to 15 kHz. This wide range, however, means that irrelevant background noise can become a problem, so that conspecific call frequencies may be adapted, even over a short time, to optimize communication. Small echo-

Amazon EX1014 Page 792

locating animals, such as bats, generally have a band of enhanced hearing sensitivity and resolution near their call frequency.

Aquatic animals, of course, operate in very different acoustic conditions to land animals, because the density of water nearly matches that of their body tissues. In addition, while the water medium may be considered to be essentially three-dimensional over small distances, it becomes effectively two-dimensional for the very long-distance communication of which some large aquatic animals are capable. Some of these matters will be discussed in Sect. 19.8.

19.2 Hearing and Sound Production

Despite the large differences in anatomy and the great separation in evolutionary time between different animals, there are many surprising similarities in their mechanisms of sound production and hearing. The basic mechanism by which sound or vibration is converted to neural sensation is usually one involving displacement of a set of tiny hairs (cilia) mounted on a cell in such a way that their displacement opens an ion channel and causes an electric potential change in the cell, ultimately leading to a nerve impulse. Such sensory hairs occur in the lateral line organs of some species of fish, supporting the otoliths in other aquatic species, and in the cochlea of land-dwelling mammals. Even the sensory cells of insects are only a little different. In the subsequent discussion there will not be space to consider these matters more fully, but they are treated in more detail in Chap. 12.

Somewhat surprisingly, the auditory sensitivities of animals differing widely in both size and anatomy vary less than one might expect. Human hearing, for example, has a threshold of about 20 μPa in its most sensitive frequency range between 500 Hz and 5 kHz, where most auditory information about conspecific communication and the cries of predators and prey is concentrated, and the sensitivity is within about 10 dB of this value over a frequency range from about 200 Hz to 7 kHz. Other mammals have very similar sensitivities with frequency ranges scaled to vary roughly as the inverse of their linear dimensions, as discussed in Sect. 19.1. Insects in general have narrower bandwidth hearing matched to song frequencies for conspecific communication. Of course, with all of these generalizations there are many outliers with very different hearing abilities.

Sound production mechanisms fall into two categories depending upon whether or not the animal is actively air-breathing. For air-breathing animals there is an internal air reservoir, the volume and pressure of which are under muscular control, so that air can be either inhaled or exhaled. When exhalation is done through some sort of valve with mechanically resonant flaps or membranes, it can set the valve into vibration, thus producing an oscillatory exhalation of air and so a sound source. In the case of aquatic mammals, which are also air-breathing, it would be wasteful to exhale the air, so that it is instead moved from one reservoir to another through the oscillating valve, the vibration of the thin walls of one of the reservoirs then radiating the sound, as is discussed in more detail in Chap. 19. Animals such as insects that do not breathe actively must generally make use of muscle-driven mechanical vibrations to produce their calls, though a few such as the cockroach *Blaberus* can actually make "hissing" noises through its respiratory system when alarmed.

The amount of muscular effort an animal is prepared to expend on vocalization varies very widely. Humans, apart from a few exceptions such as trained singers, can produce a maximum sound output of a few hundred milliwatts for a few seconds, and only about 10 mW for an extended time. In terms of body mass, this amounts to something around 0.1 mW kg^{-1}. At the other end of the scale, some species of cicada with body mass of about 1 g can produce about 1 mW of sound output at a frequency of about 3 kHz, or about 1000 mW kg^{-1} on a nearly continuous basis. An interesting comparative study of birdsongs by *Brackenbury* [19.15] showed sound powers ranging from 0.15 to 200 mW and relative powers ranging from 10 to 870 mW kg^{-1}. The clear winner on the relative power criterion was the small *Turdus philomelos* with a sound output of 60 mW and a body mass of only 69 g, equivalent to 870 mW kg^{-1}, while the loudest bird measured was the common rooster *Gallus domesticus* with a sound output during crowing of 200 mW. Its body mass of 3500 g meant, however, that its relative power was only 57 mW kg^{-1}.

Nearly all sound-production mechanisms, ranging from animals through musical instruments to electrically driven loudspeakers, are quite inefficient, with a ratio of radiated acoustic power to input power of not much more than 1%, and often a good deal less than this. In the case of air-breathing animals, the losses

are largely due to viscosity-induced turbulence above the vocal valve, while in insects that rely upon vibrating body structures internal losses in body tissue are dominant. Wide-bandwidth systems such as the human vocal apparatus are generally much less efficient than narrow-band systems, the parameters of which can be optimized to give good results at a single song frequency. To all this must then be added the internal metabolic loss in muscles, which has not been considered in the 1% figure.

19.3 Vibrational Communication

Most animals are sensitive to vibrations in materials or structures with which they have physical contact. Even humans can sense vibrations over a reasonably large frequency range by simply pressing finger tips against the vibrating object, while vibrations of lower frequency can also be felt through the feet, without requiring any specialized vibration receptors. Many insects make use of this ability to track prey and also for conspecific communication, and some for which this mode of perception is very important have specialized receptors, called subgenual organs, just below a leg joint. A brief survey for the case of insects has been given by *Ewing* [19.16].

For water gliders and other insects that hunt insects that have become trapped by surface tension on the surface of relatively still ponds, the obvious vibratory signal is the surface waves spreading out circularly from the struggles of the trapped insect. The propagation of these surface waves on water that is deep compared with the wavelength has been the subject of analysis, and the calculated wave velocity c_s, which is influenced by both the surface tension T of water and the gravitational acceleration g, is found to depend upon the wavelength λ according to the relation

$$c_s = \left(\frac{2\pi T}{\rho_w \lambda} + \frac{g\lambda}{2\pi}\right)^{1/2}, \qquad (19.1)$$

where ρ_w is the density of water. From the form of this relation it is clear that the wave speed has a minimum value when $\lambda = 2\pi(T/g\rho_w)^{1/2}$, and substituting numerical values then gives a wavelength of about 18 mm, a propagation speed of about 24 cm s^{-1}, and a frequency of about 1.3 Hz. Waves of about this wavelength will continue to move slowly across the surface after other waves have dissipated. The attenuation of all these surface waves is, however, very great.

From (19.1) it is clear that, at low frequencies and long wavelengths, the wave speed c_s is controlled by gravitational forces and $c_s \approx g/\omega$. These are called gravity waves. At high frequencies and short wavelengths, surface tension effects are dominant and $c_s \approx (T\omega/\rho_w)^{1/3}$. These are called capillary waves. The frequencies of biological interest are typically in the range 10 to 100 Hz, and so are in the in the lower part of the capillary-wave regime, a little above the crossover between these two influences, and the propagation speeds are typically in the range 30 to 50 cm s^{-1}.

For insects and other animals living aloft in trees and plants, the waves responsible for transmission of vibration are mostly bending waves in the leaves or branches. The wave speed is then proportional to the square root of the thickness of the structure and varies also as the square root of frequency. Transmission is very slow compared with the speed of sound in air, although generally much faster than the speed of surface waves on water. Once again, the attenuation with propagation distance is usually large.

Disturbances on heavy solid branches or on solid ground also generate surface waves that can be detected at a distance, but these waves propagate with nearly the speed of shear waves in the solid, and thus at speeds of order 2000 m s^{-1}. On the ground they are closely related to seismic waves, which generally have much larger amplitude, and animals can also detect these. Underground, of course, as when an animal is seeking prey in a shallow burrow, the vibration is propagated by bulk shear and compressional waves.

19.4 Insects

The variety of sound production and detection mechanisms across insect species is immense [19.3, 16–18], but their acoustic abilities and senses possess many features in common. Many, particularly immature forms such as caterpillars but also many other insects, use external sensory hairs to detect air movement and these can sense both acoustic stimuli, particularly if tuned to resonance, or the larger near-field air flows produced by the wings of predators [19.19]. These hairs respond to the viscous drag created by acoustic air flow past them and, since

this is directional, the insect may be able to gain some information about the location of the sound source.

Analysis of the behavior of sensory hairs is complex [19.19], but some general statements can be made. Since what is being sensed is an oscillatory flow of air across the surface of the insect, with the hair standing normal to this surface, it is important that the hair be long enough to protrude above the viscous boundary layer, the thickness of which varies inversely with the square root of frequency and is about 0.1 mm at 150 Hz. This means that the length of the hair should also be about inversely proportional to the frequency of the stimulus it is optimized to detect, and that it should typically be at least a few tenths of a millimeter in length. The thickness of the hair is not of immense importance, provided that it is less than about 10 μm for a typical hair, since much of the effective mass is contributed by co-moving fluid. At the mechanical resonance frequency of the hair, the sensitivity is a maximum, and the hair tip displacement is typically about twice the fluid displacement, provided the damping at the hair root is not large, but the response falls off above and below this frequency.

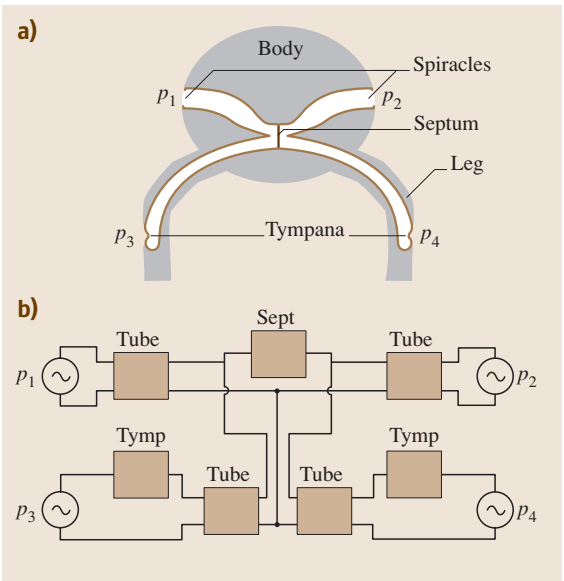


Fig. 19.2 (**a**) Schematic drawing of the auditory anatomy of a cricket. (**b**) Simplified electric network analog for this system. The impedances of the tympana and the septum are each represented by a set of components L, R, C connected in series; the tube impedances are represented by 2×2 matrices. The acoustic pressures p_i acting on the system are generally all nearly equal in magnitude, but differ in phase depending upon the direction of incidence of the sound

Mature insects of some species, however, have auditory systems that bear a superficial resemblance to those of modern humans in their overall structure. Thus for example, and referring to Fig. 19.2a, the cricket has thin sensory membranes (tympana or eardrums) on its forelegs, and these are connected by tubes (Eustachian tubes) that run to spiracles (nostrils) on its upper body, the left and right tubes interacting through a thin membrane (nasal septum) inside the body. Analysis of this system [19.9] shows that the tubes and membranes are configured so that each ear has almost cardioid response, directed ipsilaterally, at the song frequency of the insect, allowing the insect to detect the direction of sound arrival. Some simpler auditory systems showing similar directional response are discussed in Sect. 19.5.

Detailed analysis of the response of anatomically complex systems such as this is most easily carried out using electric network analogs, in which pressure is represented by electric potential and acoustic flow by electric current. Tubes are then represented as 2×2 matrices, diaphragms by L, R, C series resonant circuits, and so on, as shown in Fig. 19.2b. Brief details of this approach are given in Sect. 19.10 and fuller treatments can be found in the published literature [19.9, 20].

When it comes to sound production, insects are very different from other animals because they do not have the equivalent of lungs and a muscle-driven respiratory systems. They must therefore generally rely upon muscle-driven vibration of diaphragms somewhere on the body in order to produce sound. One significant exception is the Death's Head Hawk Moth *Acherontia atropos*, which takes in and then expels pulses of air from a cavity closed by a small flap-valve that is set into oscillation by the expelled air to produce short sound pulses [19.16].

Some insects, such as cicadas, have a large abdominal air cavity with two areas of flexible ribbed cartilage that can be made to vibrate in a pulsatory manner under muscle action, progressive buckling of the membrane as the ribs flip between different shapes effectively multiplying the frequency of excitation by as much as a factor 50 compared with the frequency of muscle contraction. The coupling between these tymbal membranes and the air cavity determines the oscillation frequency, and radiation is efficient because of the monopole nature of the source. The sound frequency is typically in the range 3 to 5 kHz, depending upon insect size, and the radiated acoustic power can be as large as 1 mW [19.21] on an almost continuous basis, though the

Amazon EX1014 Page 795

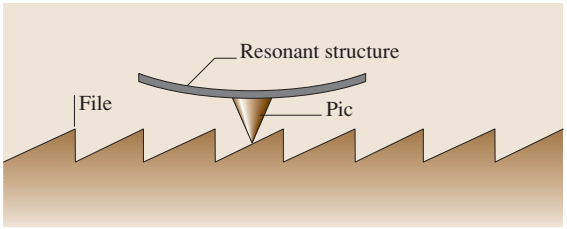


Fig. 19.3 Configuration of a typical file and pic as used by crickets and other animals to excite resonant structures on their wing cases

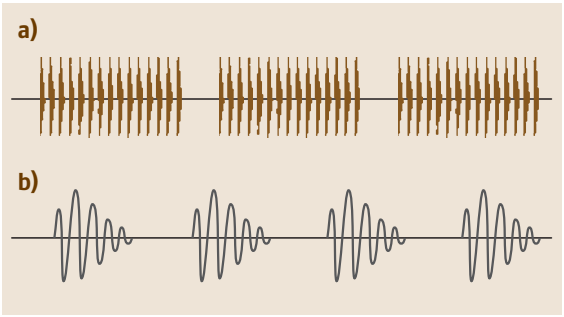


Fig. 19.4a,b Structure of the sound emitted by a typical insect: **(a)** three strokes of the file, the individual oscillations being the pic impacts on the file teeth; **(b)** expanded view of four of the pulses in **(a)**, showing resonant oscillations of the structure produced by successive pic impacts on the file

actual sound signal may consist of a series of regularly spaced bursts corresponding to tymbal collapses and rebounds. Again, there are occasional insects that have developed extreme versions of this sound producing mechanism, such as the bladder cicada *Cystosoma Saundersii (Westwood)*, in which the whole greatly extended abdomen is a resonant air sac and the song frequency is only 800 Hz.

Most other insects do not have significant bodily air cavities and must produce sound by causing parts of their body shell or wings to vibrate by drawing a ribbed file on their legs, or in some cases on their wings, across the hard rim of the membrane, as shown in Fig. 19.3. The anatomical features making up the file and vibrating structure vary very greatly across species, as has been discussed and illustrated in detail by *Dumortier* [19.22]. In some cases the file is located on the vibrating structure and the sharp pic on the part that is moving. The passage of the pic across each tooth or rib on the file generates a sharp impulse that excites a heavily damped transient oscillation of the vibrator, and each repetitive motion of the leg or other structure passes the pic over many teeth. A typical insect call can therefore be subdivided into (a) the leg motion frequency, (b) the pic impact frequency on the file, and (c) the oscillation frequency of the vibrating structure. Since the file may have 10 or more teeth, the frequency of the pic impacts will be more than 10 times the frequency of leg motion, and the structural oscillation may be 10 times this frequency. The structure of a typical call is thus of the form shown in Fig. 19.4. The dominant frequency is usually that of the structural vibration, and this varies widely from about 2 to 80 kHz depending upon species and size [19.23]. Because of the structure of the call, the frequency spectrum appears as a broad band centered on the structural vibration frequency.

Such a vibrating membrane is a dipole source and so is a much less efficient radiator than is a monopole source, except at very high frequencies where the wing dimensions become comparable with the sound wavelength excited. Small insects therefore generally have songs of higher frequency than those that are larger, in much the same way as discussed in Sect. 19.1 for other animals. Some crickets, however, have evolved the strategy of digging a horn-shaped burrow in the earth and positioning themselves at an appropriate place in the horn so as to couple their wing vibrations efficiently to the fundamental mode of the horn resonator, thus taking advantage of the dipole nature of their wing source and greatly enhancing the radiated sound power [19.24].

19.5 Land Vertebrates

The class of vertebrates that live on land is in many ways the most largely studied, since they bear the closest relation to humans, but these animals vary widely in size and behavior. Detailed discussion of birds is deferred to Sect. 19.6, since their songs warrant special attention, while human auditory and vocal abilities are not considered specifically since they are the subject of several other chapters. Reptiles, too, are largely omitted from detailed discussion, since most of them are not notably vocal, and bats are given special attention in Sect. 19.7.

The feature possessed by all the land-dwelling vertebrates is that they vocalize using air expelled from their

inflated lungs through some sort of oscillating vocal valve. The available internal air pressure depends upon the emphasis put by the species on vocalization, but is typically in the range 100–2000 Pa (or 1–20 cm water gauge) almost independent of animal size, which is what is to be expected if the breathing muscles and lung cavity diameter scale about proportionally with the linear size of the animal. There are, however, very wide variations, as discussed in Sect. 19.2. The auditory abilities of most animals are also rather similar, the 20 dB bandwidth extending over a factor of about 100 in frequency and with a center-frequency roughly inversely proportional to linear size, though again there are outliers, as shown in Fig. 19.1. Although the auditory and vocalization frequencies of different animals both follow the same trend with size, the auditory systems generally have a much larger frequency range than the vocalizations for several reasons. The first is that hearing serves many purposes in addition to conspecific communication, particularly the detection of prey and predators. The second is that, particularly with the more sophisticated animals, there are many nuances of communication, such as the vowels and consonants in human speech, for which information is coded in the upper parts of the vocalization spectrum.

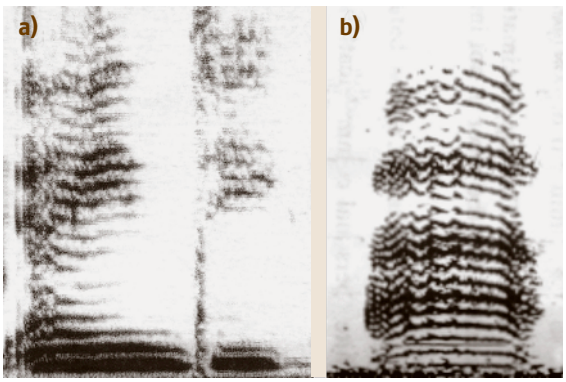


Fig. 19.5a,b A typical time-resolved spectrum for **(a)** human female speech, and **(b)** the cry of a raven, the level of each frequency component being indicated by its darkness. The frequency range is 0–6 kHz and the duration about 0.7 s in **(a)** and 0.5 s in **(b)**. It can be seen that at any instant the spectrum consists of harmonics of the fundamental, with particular bands of frequencies being emphasized. These are the formant bands that distinguish one vowel from another in humans and enable similar distinctions to be made by other animals. Consonants in human speech consist of broadband noise, an example occurring at the beginning of the second syllable in **(a)**, and other animals may have similar interpolations between the tonal sounds

For humans, for example, the primary vocalization frequency is typically in the range 100 to 300 Hz, but the frequency bands that are emphasized in vowel formants lie between 1 and 3 kHz. The same pattern is exhibited in the vocalizations of other air-breathing animals, as shown for the case of a raven in Fig. 19.5. The same formant structure is seen in the cries of elephants, with a fundamental in the range 25–30 Hz and in high-pitched bird songs, which may have a fundamental above 4 kHz.

In most such animals, sound is produced by exhaling the air stored in the lungs through a valve consisting of two taut tissue flaps or membranes that can be made to almost close the base of the vocal tract, as shown in Fig. 19.6a. The lungs in mammals are a complicated

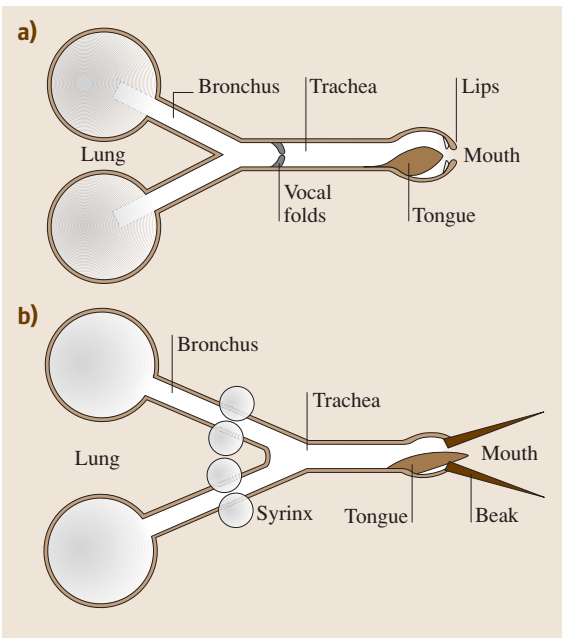


Fig. 19.6 **(a)** Sketch of the vocal system of a typical land mammal. The lungs force air through the dual-flap vocal-fold valve, producing a pulsating flow of air that is rich in harmonics of the fundamental frequency. Resonances of the upper vocal tract, which can be modified by motion of the tongue and lips, produce formants in the sound spectrum which encode vocal information. **(b)** Sketch of the vocal system of a typical song-bird. There are two inflated-membrane valves in the syrinx just below the junction of the bronchi with the trachea, which may be operated either separately or together. Again a harmonic-rich airflow is produced that can either have formants imposed by resonances of the vocal tract and beak, or can be filtered to near pure-tone form by an inflated sac in the upper vocal tract

Amazon EX1014 Page 797

quasi-fractal dendritic network of tubules with as many as 16 stages of subdivision, the final stage being terminated by alveolar sacs that provide most of storage volume. The interaction of exhaled air pressure and air flow with the flaps of this valve when they have been brought together (or adducted) causes them to vibrate at very nearly their natural frequency, as determined by mass and tension, and this in turn leads to an oscillating air flow through the valve. The classic treatment of the human vocal valve is that of *Ishizaka* and *Flanagan* [19.25], but there have been many more recent treatments exploring modifications and refinements of this model. Very similar vocal valves are found in other mammals, while the major difference in the case of birds is that the flaps of the vocal folds are replaced by opposed taut membranes inflated by air pressure in cavities behind them, and there may be two such valves, as shown in Fig. 19.6b and discussed later in Sect. 19.6.

In most cases, the vibration frequency of the valve is very much lower than the frequency of the first maximum in the upper vocal tract impedance. The valve therefore operates in a nearly autonomous manner at a frequency determined by its geometry, mass and tension and, to a less extent, by the air pressure in the animal's lungs. Because any such valve is necessarily nonlinear in its flow properties, particularly if it actually closes once in each oscillation cycle, this mechanism generates an air flow, and thus a sound wave in the vocal tract, containing all harmonics of the fundamental oscillation. The radiated amplitudes of these harmonics can be modified by resonances in the air column of the vocal tract to produce variations in sound quality and peaks in the radiated spectrum known generally as vocal formants. Changing the tension of the valve flaps changes the fundamental frequency of the vocal sound, but the frequencies of the formants can be changed independently by moving the tongue, jaws, and lips or beak.

To be more quantitative, suppose that the pressure below the valve is p_0 and that above the valve p_1. If the width of the opening is W and its oscillating dimension is $x(t)$, then the volume flow U_1 through it is determined primarily by Bernoulli's equation and is approximately

$$U_1(t) = \left(\frac{2(p_0 - p_1)}{\rho}\right)^{1/2} Wx(t)\,, \tag{19.2}$$

where ρ is the density of air. In order for the valve to be maintained in oscillation near its natural frequency f, the pressure difference $p_0 - p_1$ must vary with the flow and with a phase that is about 90° in advance of the valve opening $x(t)$. This can be achieved if the acoustic impedance of the lungs and bronchi is essentially compliant, and that of the upper vocal tract small, at the oscillation frequency. The fact that this pressure difference appears as a square root in (19.2) then introduces upper harmonic terms at frequencies $2f$, $3f, \ldots$ into the flow $U(t)$. Allowance must then be made for the fact that the vocal valve normally closes for part of each oscillation cycle, so that $x(t)$ is no longer simply sinusoidal, and this introduces further upper-harmonic terms into the flow. This is, of course, a very condensed and simplified treatment of the vocal flow dynamics, and further details can be found in the literature for the case of mammalian, and specifically human, animals [19.25] and also for birds [19.26].

This is, however, only the beginning of the analysis, for it is the acoustic flow out of the mouth that determines the radiated sound, rather than the flow from the vocal valve into the upper vocal tract. Suppose that the acoustic behavior of the upper vocal tract is represented by a 2×2 matrix, as shown in Fig. 19.11 of Sect. 19.10, and that the mouth or beak is regarded as effectively open so that the acoustic pressure p_2 at this end of the tract is almost zero. Then the acoustic volume flow U_2 out of the mouth at a particular angular frequency ω is

$$U_2 = \frac{Z_{21}}{Z_{22}} U_1\,, \tag{19.3}$$

where U_1 is the flow through the larynx or syrinx at this frequency. For the case of a simple cylindrical tube of length L at angular frequency ω, this gives

$$U_2 = \frac{U_1}{\cos kL}\,, \tag{19.4}$$

where $k = (\omega/c) - \mathrm{i}\alpha$, c is the speed of sound in air, and α is the wall damping in the vocal tract. If the spectrum of the valve flow U_1 has a simple envelope, as is normally the case, then (19.4) shows that the radiated sound power, which is proportional to $\omega^2 U_2^2$, has maxima when $\omega L/c = (2n-1)\pi/2$, and thus in a sequence $1, 3, 5, \ldots$ These are the formant bands, the precise frequency relationship of which can be varied by changing the geometry of the upper vocal tract, which in turn changes the forms of Z_{21} and Z_{22}.

Some other air-breathing animals (and even human sopranos singing in their highest register) however, adjust their vocal system so that the frequency of the vocal valve closely matches a major resonance of the upper vocal tract, usually that of lowest frequency but not necessarily so. Some species of frogs and birds achieve this by the incorporation of an inflatable sac in the upper vocal tract. In some cases the sac serves simply to provide

a resonance of appropriate frequency, and the sound is still radiated through the open mouth, but in others the walls of the sac are very thin so that they vibrate under the acoustic pressure and provide the primary radiation source. This will be discussed again in Sect. 19.6 in relation to bird song, but the analysis for frog calls is essentially the same.

The overall efficiency of sound production in vertebrates is generally only 0.1–1%, which is about the same as for musical instruments. The acoustic output power output is typically in the milliwatt range even for quite large animals, but there is a very large variation between different species in the amount of effort that is expended in vocalization.

The hearing mechanism of land-dwelling vertebrates shows considerable similarity across very many species. Sound is detected through the vibrations it induces in a light taut membrane or tympanum, and these vibrations are then communicated to a transduction organ where they are converted to nerve impulses, generally through the agency of hair cells. The mechanism of neural transduction and frequency discrimination is complex, and its explanation goes back to the days of *Helmholtz* [19.27] in the 19th century, with the generally accepted current interpretation being that of *von Békésy* [19.28] in the mid-20th century. Although these studies related specifically to human hearing, the models are generally applicable to other land-dwelling vertebrates as well, as surveyed in the volume edited by *Lewis* et al. [19.29]. All common animals have two auditory transducers, or ears, located on opposite sides of the head, so that they are able to gain information about the direction from which the sound is coming.

In the simplest such auditory system, found for example in frogs, the two ears open directly into the mouth cavity, as shown in Fig. 19.7a, one neural transducer being closely coupled to each of the two tympana. Such a system is very easily analyzed at low frequencies where the cavity presents just a simple compliance with no resonances of its own. Details of the approach are similar to those for the more complex system to be discussed next. For a typical case in which the ears are separated by 20 mm, the cavity volume is 1 cm^3, and the loaded tympanum resonance is at 500 Hz, the calculated response has the form shown in Fig. 19.7b. Directional discrimination is best at the tympanum resonance frequency and can be as much as 20 dB. In a more realistic model for the case of a frog, the nostrils must also be included, since they lead directly into the mouth and allow the entry of sound. The calculated results in this case show that the direction of maximum response for each ear is

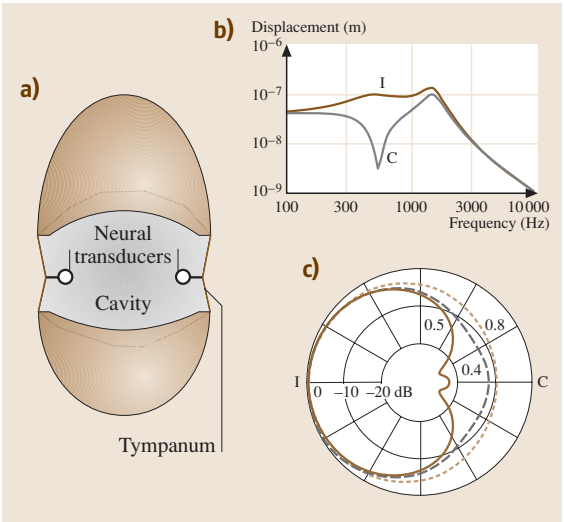


Fig. 19.7 **(a)** A simple frog-ear model; **(b)** the calculated frequency response for ipsilateral (I) and contralateral (C) sound signals for a particular optimized set of parameter values; **(c)** the calculated directivity at different frequencies, shown in kHz as a parameter (from [19.9])

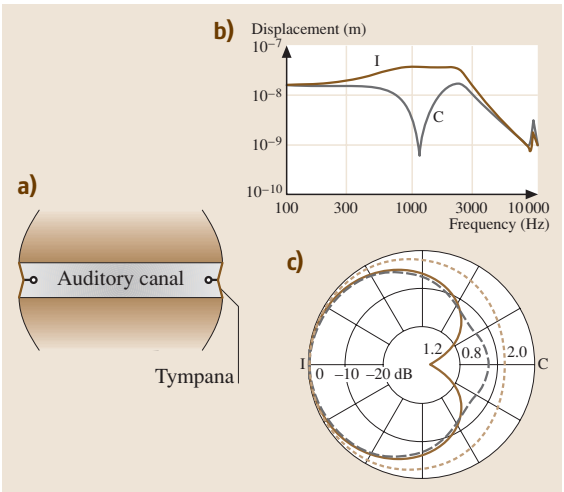


Fig. 19.8 **(a)** Idealized model of the auditory system of a lizard or bird. Each tympanum is connected to a neural transducer and the two are joined by a simple tube. **(b)** Response of a particular case of the system in **(a)** to ipsilateral (I) and contralateral (C) sound of amplitude 0.1 Pa (equivalent to 74 dB re 20 μPa). The tube diameter is assumed to be 5 mm and its length 20 mm, while the tympana have a thickness of 10 μm and are tensioned to a resonance frequency of 1000 Hz. **(c)** Directional response of this hearing system (from [19.9])

Amazon EX1014 Page 799

shifted towards the rear of the animal, typically by as much as 30° [19.9].

In another simple auditory system such as that of most reptiles, and surprisingly birds, the two tympana are located at the ends of a tube passing through the head and each is supplied with its own neural transducer, as shown in Fig. 19.8a. The tympana may be recessed at the end of short tubes to provide protection, these tubes only slightly modifying the frequency response. The behavior of such a system can be analyzed using the electric network analogs discussed in Sect. 19.10, with each tympanum involving a series L, R, C combination and the tube being represented by a 2×2 impedance matrix Z_{ij}. If the sound comes from straight in front, then it is the same in magnitude and phase at each ear, so that their responses are necessarily the same. When the sound comes from one side, however, there is a phase difference between the two ear pressures, along with a rather less significant amplitude difference. The motion of each tympanum is determined by the difference between the internal and external pressures acting upon it, and the internal pressure depends upon the signal transferred from the opposite tympanum, modified by the phase delay or resonance of the coupling tube. The analysis is straightforward, but the resulting behavior, which depends upon the mass and tension of the tympana and the length and diameter of the connecting tube, can only be determined by explicit calculation [19.9]. When these parameters are appropriately chosen, each ear has a strongly directional response near the resonance frequency of the tympana, as shown in Fig. 19.8b, c. The balance between the depth and sharpness of the contralateral minimum is affected by the acoustic absorption in the walls of the connecting canal. In some animals, the bone lining of this canal is actually porous, which reduces the effective sound-wave velocity inside it and so increases the phase shift between internal and external signals, a feature that is of assistance when the physical separation of the ears is small.

In most mammals and other large animals, the auditory system is modified in several significant ways. The first is that the auditory canal joining the two ears in birds and reptiles has generally degenerated in mammals to the extent that each ear functions nearly independently. The connecting canal in humans has become the pair of Eustachian tubes running from the middle ear cavity, which contains the bones linking the tympanum to the cochlea, down to the nasal cavities, and its main purpose is now simply to equalize internal and external static pressures and to drain excess fluids. The middle ear cavity itself is necessary in order that the enclosed air volume be large enough that it does not raise the resonance frequency of the tympanum by too much. The topology of the whole system is surprisingly similar to that of the cricket auditory system shown in Fig. 19.2a, but the functions of some of the elements are now rather different.

The other major change is that, instead of the tympana being located almost on the surface of the animal, they are buried below the surface at the end of short auditory canals (meati), which lead to external ears (pinnae) in the shape of obliquely truncated horns. As well as protecting the tympana from mechanical damage, the canals add a minor resonance of their own, generally in the upper part of the auditory range of the animal concerned. The pinnae both increase the level of the pressure signal, typically by about 10 dB and in some cases even more, and impart a directionality to the response [19.9, 30]. The convoluted form of some pinnae also imparts directionally-excited resonances that help distinguish direction from tonal cues, a feature that is necessary in order to distinguish up/down directions for sounds coming from immediately ahead or immediately behind. Some animals with particularly large pinnae, such as kangaroos, are able to rotate these independently to help locate a sound source. In all such cases, however, the neural system plays a large part in determining sound direction by comparing the phases and amplitudes, and sometimes the precise timing of transients, in the signals received by the two ears.

In a hearing system of any simple type it is clear that, if geometrical similarity is maintained and the system is simply scaled to the linear size of the animal concerned, then the frequency of maximum discrimination will vary as the inverse of the linear size of the animal, giving an approximate match to general trend of vocalization behavior. There is, however, one anomaly to be noted [19.31]. Vocalizations by some frogs have fundamental frequencies as high as 8 kHz and the vocal signal contains prominent formant bands in the case studied at 20 kHz and 80 kHz, so that they might almost be classed as ultrasonic. Despite this, behavioral evidence indicates that these frogs cannot hear signals much above 10 kHz, so that perhaps the formant bands are simply an epiphenomenon of the sound production system. The same is true of the calls of some birds.

Amazon EX1014 Page 800

19.6 Birds

Because of their generally melodious songs, birds have excited a great deal of attention among researchers, as described in several classic books [19.32, 33]. Hearing does not require further discussion because the auditory system is of the simple tube-coupled type discussed in Sect. 19.5 and illustrated in Fig. 19.8. Research interest has focused instead on the wide range of song types that are produced by different species. These range from almost pure-tone single-frequency calls produced by doves, and whistle-like sounds sweeping over more than a 2:1 frequency range produced by many other species, through wide-band calls with prominent formants, to the loud and apparently chaotic calls of the Australian sulfur-crested cockatoo. Some birds can even produce two-toned calls by using both syringeal valves simultaneously but with different tuning.

As noted in Sect. 19.5 and illustrated in Fig. 19.6b, song birds have a syrinx consisting of dual inflated-membrane valves, one in each bronchus just below their junction with the trachea. These valves can be operated simultaneously and sometimes at different frequencies, but more usually separately, and produce a pulsating air-flow rich in harmonics. In birds such as ravens, the impedance maxima of the vocal tract, as measured at the syringeal valve, produce emphasized spectral bands or formants much as in human speech, as is shown in Fig. 19.5. The vocal tract of a bird is much less flexible in shape than is that of a mammal, but the formant frequencies can still be altered by changes in tongue position and beak opening. Studies with laboratory models [19.34] show that the beak has several acoustic functions. When the beak is nearly closed, it presents an end-correction to the trachea that is about half the beak length at high frequencies, but only about half this amount at low frequencies, so that the formant resonances are brought closer together in frequency. As the beak gape is widened, the end-correction reduces towards the normal value for a simple open tube at all frequencies. The beak also improves radiation efficiency, particularly at higher frequencies, by about 10 dB for the typical beak modeled. Finally, the beak enhances the directionality of the radiated sound, particularly at high frequencies and for wide gapes, this enhanced directionality being as much as 10 dB compared with an isotropic radiator.

The role of the tongue has received little attention as yet, partly because it is rather stiff in most species of birds. It does, however, tend to constrict the passage between the top of the trachea and the root of the beak, and this constriction can be altered simply by raising the tongue, thereby exaggerating the end-correction provided to the trachea by the beak. These models of beak behavior can be combined with models of the syrinx and upper vocal tract [19.26] to produce a fairly good understanding of vocalization in birds such as ravens. The fundamental frequency of the song can be changed by changing the muscle tension and thus the natural resonance frequency of the syringeal valve, while the formants, and thus the tone of the song, can be varied by changing the beak opening and tongue position. The role of particular anatomical features and muscular activities in controlling the fundamental pitch of the song has since been verified by careful measurement and modeling [19.35, 36].

Some birds have developed the ability to mimic others around them, a notable example being Australian Lyrebirds of the family *Menuridae* which, as well as imitating other birds, have even learned to mimic human-generated sounds such as axe blows, train whistles, and even the sound of film being wound in cameras. There has also been considerable interest in the vocalization of certain parrots and cockatoos, which can produce, as well as melodious songs, quite intelligible imitations of human speech. The important thing here is to produce a sound with large harmonic content and then to tune the vocal tract resonances, particularly those that have frequencies in the range 1–2.5 kHz, to match those of the second and third resonances of the human vocal tract, which are responsible for differentiating vowel sounds. Consonants, of course, are wide-band noise-like transients. Careful studies of various parrots [19.37, 38] show that they can achieve this by careful positioning of the tongue, much in the way that humans achieve the same thing but with quantitative differences because of different vocal tract size. The match to human vowels in the second and third human formants can be quite good. The first formant, around 500 Hz in humans, is not reproduced but this is not important for vowel recognition.

These studies produce a good understanding of the vocalization of normal song-birds, the calls of which consist of a dominant fundamental accompanied by a range of upper harmonics. Less work has been done on understanding the nearly pure-tone cries of other birds. For many years there has been the supposition that the sound production mechanism in this case was quite different, perhaps involving some sort of aerodynamic whistle rather than a syringeal vibration. More recent experimental observations on actual singing birds, using

modern technology [19.39], have established however that, at least in the cases studied, the syringeal valve is in fact normally active, though it does not close completely in any part of the cycle, thus avoiding the generation of large amplitudes of upper harmonics. Suppression of upper harmonics can also be aided by filtering in the upper vocal tract [19.40].

In the case of doves, which produce a brief pure-tone "coo" sound at a single frequency, the explanation appears to lie in the use of a thin-walled inflated esophageal sac and a closed beak, with fine tuning provided by an esophageal constriction [19.41]. The volume V of air enclosed in the sac provides an acoustic compliance $C = V/\rho c^2$ that is in parallel with an acoustic inertance $L = m/S^2$ provided by the mass m and surface area S of the sac walls. The resonance frequency f of this sac filter is given by

$$f = \frac{1}{2\pi}\left(\frac{\rho c^2 S^2}{mV}\right)^{1/2}, \tag{19.5}$$

where ρ is the density of air and c is the speed of sound in air. When excited at this frequency by an inflow of air from the trachea, the very thin expanded walls of the sac vibrate and radiate sound to the surrounding air. For a typical dove using this strategy, the wall-mass and dimensions of the inflated sac are about right to couple to the "coo" frequency of about 900 Hz. From (19.5), the resonance frequency varies only as the square root of the diameter of the sac, so that a moderate exhalation of breath can be accommodated without much disturbance to the resonance. The dove must, however, learn to adjust the glottal constriction to couple the tracheal resonance efficiently to that of the sac.

Some other birds, such as the Cardinal, that produce nearly pure-tone calls over a wide frequency range, appear to do so with the help of an inflated sac in the upper vocal tract that leads to the opened beak, with the tracheal length, sac volume, glottal constriction, tongue position, and beak opening all contributing to determine the variable resonance frequency of the filter. Because the sac is not exposed as in the dove, its walls are heavy and do not vibrate significantly. It thus acts as a Helmholtz resonator, excited by the input flow from the trachea and vented through the beak. The bird presumably learns to adjust the anatomical variables mentioned above to the syringeal vibration frequency, which is in turn controlled by syringeal membrane tension, and can then produce a nearly pure-tone song over quite a large frequency range. Despite this explanation of the mechanism of pure-tone generation in some cases, others have still to be understood, for the variety of birdsong is so great that variant anatomies and vocalization skills may well have evolved.

With most of the basic acoustic principles of birdsong understood, most of the interest in birdsong centers upon the individual differences between species. The variety found in nature is extremely great and, while sometimes correlated with large variations in bird anatomy, it also has many environmental influences. Some birds, surprisingly, have songs with prominent formant bands extending well into the ultrasonic region, despite the fact that behavioral studies and even auditory brainstem measurements show that they have no response above about 8 kHz [19.31, 42]. The formant bands in these cases are therefore presumably just an incidental product of the vocalization mechanism.

19.7 Bats

Since human hearing, even in young people, is limited to the range from about 20 Hz to about 20 kHz, frequencies lying outside this range are referred to as either infrasonic or ultrasonic. Very large animals such as elephants may produce sounds with appreciable infrasonic components, though the higher harmonics of these sounds are quite audible. Small animals such as bats, however, produce echo-location calls with dominant frequencies around 80 kHz that are inaudible to humans, though some associated components of lower frequency may be heard. In this section brief attention will be given to the sonar abilities of bats, though some other animals also make use of ultrasonic techniques [19.43].

Sound waves are scattered by any inhomogeneities in the medium through which they are propagating, the density of an animal body in air provides a large contrast in acoustic impedance and therefore a large scattered signal. The basic physics shows that, for a target in the form of a sphere of diameter d and a signal of wavelength $\lambda \gg d$, the echo strength is proportional to d^6/λ^4 times the incident intensity. This incident intensity is itself determined by the initial signal power and the angular width of the scanning signal, which varies about as λ/a, where a is the radius of the mouth of the probing animal. If R is the distance between the probing animal and the target, then the sound power incident on the target varies as $a^2/\lambda^2 R^2$ and the returned echo strength E thus varies

as

$$E \propto \frac{d^6 a^2}{\lambda^6 R^4} , \tag{19.6}$$

provided the wavelength λ is much greater than the diameter d of the scattering object and the radius a of the mouth. Another factor must, however, be added to this equation and that is the attenuation of the sound due to atmospheric absorption. This attenuation depends on relative humidity, but is about 0.1 dB m^{-1} at 10 kHz, increasing to about 0.3 dB m^{-1} at 100 kHz, and so is not a very significant factor at the short ranges involved.

Once the size of the target becomes comparable with the incident sound wavelength, the scattering tends towards becoming reflection and geometrical-optics techniques can be used. In this limit, the reflected intensity is proportional to the area of the target, but depends in a complicated way upon its shape and orientation. Finally, if a pure-tone sound is used, any Doppler shift in the frequency of the echo can give information about the relative motion of prey and predator. The reflected sound can therefore give information about the nature of the target and its motion, both of which are important to the pursuing bat.

The sonar-location cries of bats and other animals are, however, not pure-tone continuous signals but rather sound pulses with a typical frequency in the range 40–80 kHz and so a wavelength of about 4–8 mm. This wavelength is comparable to the size of typical prey insects so that the echo can give information about size, shape and wing motion, while the echo delay reveals the distance. The bat's large ears are also quite directional at these frequencies, so that further information on location is obtained. Rather than using short bursts of pure tone, some predators use longer bursts in which the frequency is swept rapidly through a range of several kilohertz. This technique is known as "chirping", and it allows the returned echo to be reconstructed as a single pulse by frequency-dependent time delay, although animal neural systems may not actually do this. The advantage of chirping is that the total energy contained in the probing pulse, and thus the echo, can be made much larger without requiring a large peak power.

Because of the high frequencies involved in such sonar detection, bats have a hearing system that is specialized to cover this high-frequency range and to provide extended frequency resolution near the signal frequency, perhaps to aid in the neural equivalent of "dechirping" the signal and detecting Doppler frequency shifts. The auditory sensitivity curve of bats is similar in shape to that of humans (threshold about 20 μPa), but shifted up in frequency by a factor of about 20 to give high sensitivity in the range 10–100 kHz [19.1]. The threshold curve generally shows a particular sensitivity maximum near the frequency of the echo-location call, as might be expected. Bats also generally have forward-facing ears that are large in comparison with their body size. The functions of these in terms of increased directionality and larger receiving area are clear [19.30].

Sound production in bats is generally similar to that in other animals, but with the system dimensions tailored to the higher frequencies involved, and often with the sound emitted through the nose rather than the mouth. The complicated geometry of the nasal tract, which has several spaced enlargements along its length, appears to act as an acoustic filter which suppresses the fundamental and third harmonic of the emitted sound while reinforcing the second harmonic, at least in some species [19.44]. The physics of this arrangement is very similar to that of matching stubs applied to high-frequency transmission lines.

19.8 Aquatic Animals

The generation and propagation of sound under water, as discussed in Chap. 19, have many features that are different from those applying to sound in air. From a physical point of view, a major difference is the fact that water is often quite shallow compared with the range of sound, so that propagation tends to be two-dimensional at long distances. From the present biological viewpoint the major difference is that the acoustic impedance of biological tissue is nearly equal to that of the surrounding water, while in air these differ by a factor of about 3500. This leads to some very significant differences in the auditory anatomy and behavior between aquatic and air-dwelling animals.

The variety of aquatic animal species is comparable to that of land-dwelling animals, and they can be divided into three main categories for the purposes of the present chapter. First come the crustaceans, such as shrimp, lobsters and crabs, which are analogous to insects, then the wide variety of fish with backbones, and finally the air-breathing aquatic mammals such as the

cetaceans (whales, dolphins and porpoises, which are the subject of Chap. 19), though animals such as turtles also come into this category. There are, of course, many other categories of marine animal life, but they are not generally significant from an acoustic point of view. A good review has been given by *Hawkins* and *Myrberg* [19.45].

The crustaceans, like most insects, generally produce sound by rubbing a toothed leg-file against one of the plates covering their body, thus producing a pulse of rhythmic clicks. Some animals, such as shrimp, may produce a single loud click or pop by snapping their claws in much the same way that a human can do with fingers. Because this activity takes place under water, which has density ρ about 800 times greater and sound velocity c about 4.4 times greater than that of air, and thus a specific acoustic impedance ρc about 3500 times greater than air, the displacement amplitude of the vibration does not need to be large, and the vibrating structure is generally small and stiff. The peak radiated acoustic power in the "click" can, however, be very large – as much as several watts in the Pistol shrimp *Alpheidae* – probably from a mechanism involving acoustic cavitation. The click lasts for only a very short time, however, so that the total radiated energy is small.

Fish species that lack an air-filled "swim-bladder", the main purpose of which is to provide buoyancy, must produce sound in a similar manner. For fish that do have a swim-bladder, however, the sound production process is much more like that of the cicada, with some sort of plate or membrane over the bladder, or even the bladder wall itself, being set into oscillation by muscular effort and its displacement being aided by the compressibility of the air contained in the bladder. The surrounding tissue and water provide a very substantial mass loading which also contributes to determining the primary resonance frequency of the bladder. This is a much more efficient sound production process than that of the crustaceans and leads to much louder sounds, again generally consisting of a train of repetitive pulses at a frequency typically in the range 100–1000 Hz.

When it comes to sound detection, hair cells again come into play in a variety of ways. Because the density of the animal body is very similar to that of the surrounding water, a plane wave of sound tends to simply move the animal and the water in just the same way. The slight difference in density, however, does allow for some differential motion. The animal body is, however, relatively rigid, so that its response to diverging waves in the near field is rather different from that of the liquid. Any such relative motion between the body and the surrounding water can be detected by light structures such as hairs protruding from the body of the animal, and these can form the basis of hair-cell acoustic detectors. Even more than this, since such cells are generally sensitive to hair deflection in just one direction, arrays of hair cells can be structured so as to give information about the direction of origin of the sound. Although some such detector hairs may be mechanically strong and protrude from the body of the animal, others may be very small and delicate and may be protected from damage by being located in trench-like structures, these structures being called lateral-line organs. Their geometry also has an effect on sound localization sensitivity.

There is another type of hair detector that responds to acceleration rather than relative displacement and that therefore takes advantage of the fact that the body of the animal tends to follow the acoustic displacement of the surrounding water. This is the otolith, which consists of a small dense calcareous stone supported by one or more hair cells,as shown in Fig. 19.9. Since the motion of this stone tends to lag behind the motion of the liquid, and thus of its supports, the supporting hairs are deflected in synchrony with the acoustic stimulus, so that it can be detected by the cells to which they are attached. These sensors will normally have evolved so that combined mass of the otolith and stiffness of the hair give a resonance frequency in the range of interest to the animal, and detection sensitivity will be greatest at this frequency.

If the fish possesses an air-filled swim-bladder, then this can also be incorporated in the hearing sys-

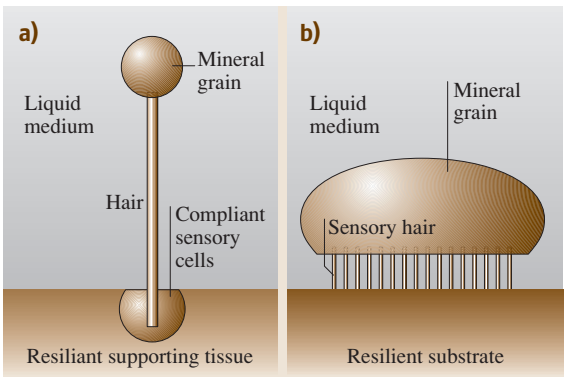


Fig. 19.9a,b Two possible forms for a simple otolith detector: **(a)** a single-cell detector, **(b)** a multi-cell detector. As the mineral grain moves sideways relative to the substrate, the sensory hairs are deflected and open ion channels in the supporting transducer cells (from [19.9])

tem [19.46]. An air bubble in water is made to move in a sound wave by the differential pressure across it. Because its mass is small and the pressure gradient is the same as that on an equivalent volume of water, the bubble tends to move more than the surrounding liquid but, in order to do so, it must displace a significant amount of this liquid. The final result is that the amplitude of free bubble motion is about three times the amplitude of the motion of the surrounding liquid. Since the body of the fish tends to follow the liquid motion because its density is nearly the same, there is relative motion between the bladder and the rest of the fish, provided it is not rigidly enclosed. This relative motion can be detected by neural cells pressed against the bladder wall and provides an efficient sensor of underwater sound. The radial compressive resonances of the swim bladder can also be important in the detection of sound and, since these are also involved in sound production in these fish, there is a good match between call frequency and the frequency of maximum hearing sensitivity, which typically lies in the range 100 to 1000 Hz. Fish with swim bladders have thus evolved to use these primarily buoyancy-regulating structures for both the production and detection of sound.

19.9 Generalities

This brief survey has shown the strong resemblance between the basic sound-producing and auditory mechanisms in a wide variety of animals. Certainly there is a distinction in the matter of sound production between those animals that store air within their bodies and those that do not, and a rather similar distinction in the matter of sound detection, but the similarities are more striking than the differences. In particular, the universal importance of hair-cells in effecting the transduction from acoustic to neural information is most striking.

Also notable in its generality is the scaling of the dominant frequency of produced sound and the associated frequency of greatest hearing sensitivity approximately inversely with the linear dimensions of the animal involved. Certainly there are notable outliers in this scaling, but that in itself is a clue to the significantly different acoustic behavior of the animal species concerned.

Fortunately, quite straightforward acoustic analysis provides a good semi-quantitative description of the operation of most of these animal acoustic systems, and in some cases such analysis can reveal evolutionarily optimized solutions to modern acoustical design problems.

19.10 Quantitative System Analysis

Biological systems are almost always anatomically complex, and this is certainly true of vocal and auditory systems, which typically comprise sets of interconnecting tubes, horns, cavities and membranes driven over a wide frequency range by some sort of internal or external excitation. While a qualitative description of the way in which such a system works can be helpful, this must be largely based upon conjectures unless one is able to produce a quantitative model for its behavior. Fortunately a method has been developed for the construction of detailed analytical models for such systems from which their acoustic behavior can be quantitatively predicted, thus allowing detailed comparison with experiment. This section will outline how this can be done, more detail being given in [19.9] and [19.20].

In a simple mechanical system consisting of masses and springs, the quantities of interest are the force F_i applied at point i and the resulting velocity of motion v_j at point j, and we can define the mechanical impedance Z^{mech} by the equation $Z^{\text{mech}}_{jj} = F_j/v_j$. Note that the subscripts on F and v are here both the same, so that Z^{mech}_{jj} is the ratio of these two quantities measured at a single point j. It is also possible to define a quantity $Z^{\text{mech}}_{ij} = F_i/v_j$ which is called the transfer impedance between the two points i and j. There is an important theorem called the reciprocity theorem, which shows that $Z_{ji} = Z_{ij}$.

In such a simple mechanical system, as well as the masses and springs, there is normally some sort of loss mechanism, usually associated with internal viscous-type losses in the springs, and the situation of interest involves the application of an oscillating force of magnitude F and angular frequency ω. We denote this force in complex notation as $F\,\mathrm{e}^{\mathrm{i}\omega t}$ where $\mathrm{i} = \sqrt{-1}$. (Note

that, in the interest of consistency throughout this Handbook, this notation differs from the standard electrical engineering notation in which i is replaced by j.) For the motion of a simple mass m under the influence of a force F we know that $F = m\,\mathrm{d}v/\mathrm{d}t$ or $F = \mathrm{i}\omega m v$, so that the impedance provided by the mass is $\mathrm{i}\omega m$. Similarly, for a spring of compliance C, $F = \int v\,\mathrm{d}t/C = -\mathrm{i}/(\omega C)$, giving an impedance $-\mathrm{i}/(\omega C)$. For a viscous loss within the spring there is an equivalent resistance R proportional to the speed of motion, so that $F = Rv$. Examination of these equations shows that they are closely similar to those for electric circuits if we assume that voltage is the electrical analog of force and current the analog of velocity. The analog of a mass is thus an inductance of magnitude m, the analog of a compliant spring is a capacitance C, and the analog of a viscous resistance is an electrical resistance R. A frictional resistance is much more complicated to model since its magnitude is constant once sliding occurs but changes direction with the direction of motion, so that its whole behavior is nonlinear.

These mechanical elements can then be combined into a simple circuit, two examples being shown in Fig. 19.10. In example (a) it can be seen that the motion of the mass, the top of the spring, and the viscous drag are all the same, so that in the electrical analog the same current must flow through each, implying a simple series circuit as in (b). The total input impedance is therefore the sum of the three component impedances, so that

$$Z^{\mathrm{mech}} = R + \mathrm{i}\omega m - \frac{\mathrm{i}}{\omega C} = R + \frac{\mathrm{i}m}{\omega}\left(\omega^2 - \frac{1}{mC}\right), \tag{19.7}$$

which has an impedance minimum, and thus maximum motion, at the resonance frequency $\omega^* = 1/(mC)^{1/2}$. In example (c), the displacement of the top of the spring is the sum of the compressive displacement of the spring and the motion of the suspended mass, and so the analog current is split between the spring on the one hand and the mass on the other, the viscous drag having the same motion as the spring, leading to the parallel resonant circuit shown in (d). There is an impedance maximum, and thus minimum motion of the top of the spring, at the resonance frequency, which is again $\omega^* = 1/(mC)^{1/2}$. In a simple variant of these models, the spring could be undamped and the mass immersed in a viscous fluid, in which case the resistance R would be in series with the inductance rather than the capacitance. These analogs can be extended to include things such as levers, which appear as electrical transformers with turns ratio equal to the length ratio of the lever arms.

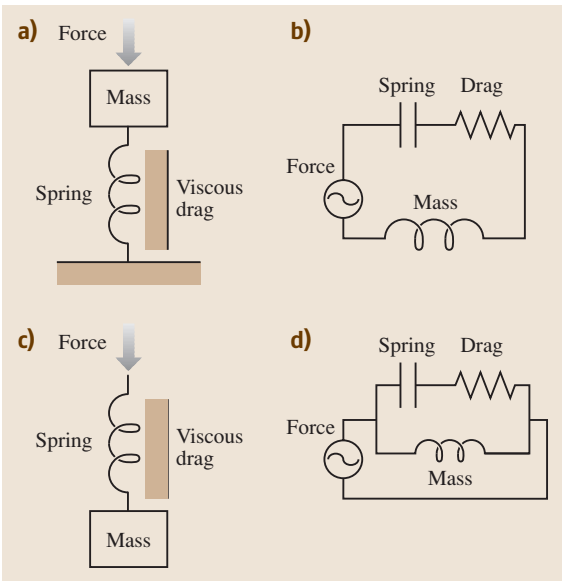


Fig. 19.10 **(a)** A simple resonant assembly with an oscillating force applied to a mass that is resting on a rigidly supported spring, the spring having internal viscous losses. **(b)** Analog electric network for the system shown in **(a)**. **(c)** An alternative assembly with the mass being freely suspended from the spring and the oscillating force being applied to the free end of the spring. **(d)** Analog electric network for the system shown in **(c)**. (Note the potential source of confusion in that an electrical inductance symbol looks like a physical spring but actually represents a mass)

While such combinations of mechanical elements are important in some aspects of animal acoustics, such as the motion of otolith detectors, most of the systems of interest involve the generation and detection of acoustic waves rather than of mechanical vibrations. The important physical variables are then the acoustic pressure p and the acoustic volume flow U, both being functions of the frequency ω. By analogy with mechanical systems, the acoustic impedance is now defined as $Z^{\mathrm{acoust}} = p/U$. For a system with a well-defined measurement aperture of area S, it is clear that $p = F/S$ and $U = vS$, so that $Z^{\mathrm{acoust}} = Z^{\mathrm{mech}}/S^2$, and once again we can define both the impedance at an aperture and the transfer impedance between two apertures. In what follows, the superscript "acoust" will be omitted for simplicity.

Consider first the case of a circular diaphragm under elastic tension and excited by an acoustic pressure $p\mathrm{e}^{\mathrm{i}\omega t}$ on one side. This problem can easily be solved using a simple differential equation [19.9], but here we seek a network analog. The pressure will deflect the dia-

Amazon EX1014 Page 806

phragm into a nearly spherical-cap shape, displacing an air volume as it does so. If the small mass of the deflected air is neglected, then deflection of the diaphragm will be opposed by its own mass inertia, by the tension force and by internal losses. The mass contributes an acoustic inertance (or inductance in the network analog) with magnitude about $2\rho_s d/S$, where ρ_s is the density of the membrane material, d is the membrane thickness, and S is its area, while the tension force T contributes an equivalent capacitance of magnitude about $2S^2/T$. This predicts a resonance frequency $\omega^* \approx 0.5(T/S\rho_s d)^{1/2}$, while the rigorously derived frequency for the first mode of the diaphragm replaces the 0.5 with 0.38. Because biological membranes are not uniform in thickness and are usually loaded by neural transducers, however, it is not appropriate to worry unduly about the exact values of numerical factors. For a membrane there are also higher modes with nodal diameters and nodal circles, but these are of almost negligible importance in biological systems.

The next thing to consider is the behavior of cavities, apertures, and short tubes. The air enclosed in a rigid container of volume V can be compressed by the addition of a further small volume $\mathrm{d}V$ of air and the pressure will then rise by an amount $p_0\gamma\,\mathrm{d}V/V$ where p_0 is atmospheric pressure and $\gamma = 1.4$ is the ratio of specific heats of air at constant pressure and constant volume. The electric analog for this case is a capacitance of value $\gamma p_0/V = \rho c^2/V$ where c is the speed of sound in air and ρ is the density of air.

In the case of a tube of length l and cross-section S, both very much less than the wavelength of the sound, the air within the tube behaves like a mass ρlS that is set into motion by the pressure difference p across it. The acoustic inertance, or analog inductance, is then $Z = \rho l/S$. For a very short tube, such as a simple aperture, the motion induced in the air just outside the two ends must be considered as well, and this adds an "end-correction" of 0.6–0.8 times the tube radius [19.47] or about $0.4S^{1/2}$ to each end of the tube, thus increasing its effective length by twice this amount.

These elements can be combined simply by considering the fact that acoustic flow must be continuous and conserved. Thus a Helmholtz resonator, which consists of a rigid enclosure with a simple aperture, is modeled as a capacitor and an inductor in series, with some resistance added due to the viscosity of the air moving through the aperture and to a less extent to radiation resistance. The analog circuit is thus just as in Fig. 19.10b and leads to a simple resonance.

If any of the components in the system under analysis is not negligibly small compared with the wavelength of the sound involved, then it is necessary to use a more complex analysis. For the case of a uniform tube of length l and cross section S, the pressure and flow at one end can be denoted by p_1 and U_1 respectively and those at the other end by p_2 and U_2. The analog impedance is then a four-terminal element, as shown in Fig. 19.11. The equations describing the behavior of this element are

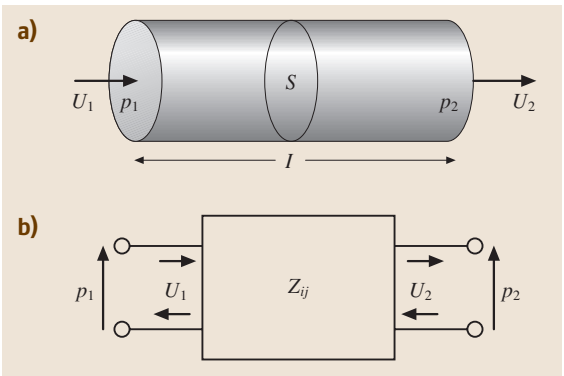


Fig. 19.11 **(a)** A simple uniform tube showing the acoustic pressures p_i and acoustic volume flows U_i. **(b)** The electrical analog impedance element Z_{ij}. Note that in both cases the current flow has been assumed to be antisymmetric rather than symmetric

$$p_1 = Z_{11}U_1 - Z_{12}U_2\,,$$
$$p_2 = Z_{21}U_1 - Z_{22}U_2\,. \qquad (19.8)$$

Note that the acoustic flow is taken to be inwards at the first aperture and outwards at the second, and this asymmetry gives rise to the minus signs in (19.8). Some texts take both currents to be inflowing, which removes the minus signs. For a simple uniform tube, symmetry demands that $Z_{21} = Z_{12}$ and that $Z_{22} = Z_{11}$. The first of these relations, known as reciprocity, is universal and does not depend upon symmetry, while the second is not true for asymmetric tubes or horns.

For a uniform tube of length l and cross-section S, consideration of wave propagation between the two ends shows that

$$Z_{11} = Z_{22} = -\mathrm{i}Z_0 \cot kl\,,$$
$$Z_{21} = Z_{12} = -\mathrm{i}Z_0 \csc kl\,, \qquad (19.9)$$

where $Z_0 = \rho c/S$ is the characteristic impedance of the tube. The propagation number k is given by

$$k = \frac{\omega}{c} - \mathrm{i}\alpha \qquad (19.10)$$

where α represents the attenuation caused by viscous and thermal losses to the tube walls, and has the approximate value

$$\alpha \approx 10^{-5}\frac{\omega^{1/2}}{a} \qquad (19.11)$$

where a is the tube radius. Since auditory systems often involve quite narrow tubes, this term can become very important.

Similar but much more complicated relations apply for horns of various shapes, the simplest being conical, exponential or parabolic in profile [19.9]. As discussed elsewhere [19.47], horns act as acoustic transformers, basically because $Z_{22} \neq Z_{11}$, and can raise the acoustic pressure in the narrow throat by about the ratio of the mouth diameter to the throat diameter. This amplification is limited, however, to frequencies high enough that the horn diameter does not change greatly over an axial distance of about one wavelength.

With the aid of these electric circuit element analogs it is now possible to construct an analog network to represent the whole acoustic system, as shown by the example in Fig. 19.2. The major guiding principle in constructing such a network is the conservation of acoustic flow, or equivalently of electric current. In the case of an auditory system, there may be several openings connecting to the environment, and the acoustic pressures at these openings will generally differ in both phase and amplitude, depending upon the direction from which the sound is coming. Phase differences are generally more important for directional discrimination than are amplitude differences, particularly for small animals, just as in directional microphones, [19.47] and the combination of tubes, cavities and membranes in the system provides additional phase shifts. The sensory input is normally generated by the motion of tympanic membranes on the two sides of the animal, with the combination of external and internal pressure determining the motion of each membrane, as shown for example in Figs. 19.7 and 19.8.

When sound production is considered, it is usual to make the approximation of separating the basic sound-generation mechanism, such as the air flow through vibrating vocal folds, from the acoustics of the subsequent tubes and cavities, thus leading to a "source/filter model". With this simplification, the source can be assumed to produce an oscillating acoustic flow at a particular frequency ω, generally accompanied by a set of harmonics at frequencies $n\omega$ generated by nonlinearity in the vibration of the flow valve. The sound production at each frequency is then determined by the final acoustic flow $U(n\omega)$ through an aperture into the surroundings. The square of this flow, multiplied the acoustic radiation resistance of the aperture, which depends upon its size, [19.47] then determines the radiated acoustic power at that frequency. In a much more complex model, the pressure generated by the acoustic input flow can be recognized to act back upon the acoustic source itself, thus modifying its behavior. Typically this results in a small shift in the source frequency to align it more closely with a neighboring strong resonance of the filter. Such interactions are well known in other systems such as musical wind instruments and probably assist in coupling the oscillations of the syrinx to the resonance of the vocal filter in "pure tone" birdsong.

Finally, it is worthwhile to emphasize that biological acoustic systems do not consist of straight tubes of uniform diameter with rigid walls, unloaded circular membranes, or other idealized elements. While it is possible in principle to include such refinements in a computational model, this will generally make that model so complex to analyze that it reveals little about the actual operation of the system. It is usually better, therefore, to construct the simplest possible model that includes all the relevant acoustic elements and to examine its predictions and the effect that changing the dimensions of some of those elements has on the overall behavior. In this way the behavior of the real biological system can be largely understood.

Remark

The number of books and papers published in this field is very large, so that the reference list below is a rather eclectic selection. Generally descriptive and behavioral references have not been included, and the list concentrates on those that apply physical and mathematical analysis to bioacoustic systems.

References

19.1 W.C. Stebbins: *The Acoustic Sense of Animals* (Harvard Univ. Press, Cambridge 1983)

19.2 J.W. Bradbury, S.L. Vehrenkamp: *Principles of Animal Communication* (Sinauer, Sunderland 1998)

19.3 R.-G. Busnel (Ed.): *Acoustic Behavior of Animals* (Elsevier, New York 1963)

19.4 A.N. Popper, R.R. Fay (Eds.): *Comparative Studies of Hearing in Vertebrates* (Springer, Berlin, New York 1980)

Amazon EX1014 Page 808

19.5 B. Lewis (Ed.): *Bioacoustics: A Comparative Approach* (Academic, London 1983)
19.6 D.B. Webster, R.R. Fay, A.N. Popper (Eds.): *The Evolutionary Biology of Hearing* (Springer, Berlin, New York 1992)
19.7 R.R. Fay, A.N. Popper (Eds.): *Comparative Hearing: Mammals* (Springer, Berlin, New York 1994)
19.8 R.R. Hoy, A.N. Popper, R.R. Fay (Eds.): *Comparative Hearing: Insects* (Springer, Berlin, New York 1998)
19.9 N.H. Fletcher: *Acoustic Systems in Biology* (Oxford Univ. Press, New York 1992)
19.10 D.A.W. Thompson: *On Growth and Form (abridged)*, 2nd edn., ed. by J.T. Bonner (Cambridge Univ. Press, Cambridge 1961)
19.11 G.B. West, J.H. Brown: Life's universal scaling laws, Phys. Today **57**(9), 36–42 (2004)
19.12 N.H. Fletcher: A simple frequency-scaling rule for animal communication, J. Acoust. Soc. Am. **115**, 2334–2338 (2004)
19.13 M. Garstang, D. Larom, R.Raspet, M. Lindeque: Atmospheric controls on elephant communication, J. Exper. Biol. **198**, 939–951 (1995)
19.14 K. McComb, D. Reby, L. Baker, C. Moss, S. Sayialel: Long-distance communication of acoustic cues to social identity in African elephants, Anim. Behav. **65**, 317–329 (2003)
19.15 J.H. Brackenbury: Power capabilities of the avian sound-producing system, J. Exp. Biol. **78**, 163–166 (1979)
19.16 A.W. Ewing: *Arthropod Bioacoustics: Neurobiology and Behavior* (Cornell Univ. Press, Ithaca 1989)
19.17 A. Michelsen, H. Nocke: Biophysical aspects of sound communication in insects, Adv. Insect Physiol. **10**, 247–296 (1974)
19.18 K. Kalmring, N. Elsner (Eds.): *Acoustic and Vibrational Communication in Insects* (Parey, Berlin, Hamburg 1985)
19.19 N.H. Fletcher: Acoustical response of hair receptors in insects, J. Comp. Physiol. **127**, 185–189 (1978)
19.20 N.H. Fletcher, S. Thwaites: Physical models for the analysis of acoustical systems in biology, Quart. Rev. Biophys. **12**(1), 25–65 (1979)
19.21 H.C. Bennet-Clark: Resonators in insect sound production: how insects produce loud pure-tone songs, J. Exp. Biol. **202**, 3347–3357 (1999)
19.22 B. Dumortier: Morphology of sound emission apparatus in arthropoda. In: *Acoustic Behaviour of Animals*, ed. by R.-G. Busnel (Elsevier, Amsterdam 1963) pp. 277–345
19.23 B. Dumortier: The physical characteristics of sound emission in arthropoda. In: *Acoustic Behaviour of Animals*, ed. by R.-G. Busnel (Elsevier, Amsterdam 1963) pp. 346–373
19.24 A.G. Daws, H.C. Bennet-Clark, N.H. Fletcher: The mechanism of tuning of the mole cricket singing burrow, Bioacoust. **7**, 81–117 (1996)
19.25 K. Ishizaka, J.L. Flanagan: Synthesis of voiced sounds from a two-mass model of the vocal cords, Bell Syst. Tech. J. **51**(6), 1233–1268 (1972)
19.26 N.H. Fletcher: Bird song – a quantitative acoustic model, J. Theor. Biol. **135**, 455–481 (1988)
19.27 H.L.F. Helmholtz: *On the Sensations of Tone as a Physiological Basis for the Theory of Music* (Dover, New York 1954), trans. A.J. Ellis
19.28 G. von Békésy: *Experiments in Hearing* (McGraw-Hill, New York 1960)
19.29 E.R. Lewis, E.L. Leverenz, W.S. Bialek: *The Vertebrate Inner Ear* (CRC, Boca Raton 1985)
19.30 N.H. Fletcher, S. Thwaites: Obliquely truncated simple horns: idealized models for vertebrate pinnae, Acustica **65**, 194–204 (1988)
19.31 P.M. Narins, A.S. Feng, W. Lin, H.-U. Schnitzler, A. Densinger, R.A. Suthers, C.-H. Xu: Old world frog and bird vocalizations contain prominent ultrasonic harmonics, J. Acoust. Soc. Am. **115**, 910–913 (2004)
19.32 C.H. Greenewalt: *Bird Song: Acoustics and Physiology* (Smithsonian Inst. Press, Washington 1968)
19.33 D.E. Kroodsma, E.H. Miller (Eds.): *Acoustic Communication in Birds*, Vol. 1-2 (Academic, New York 1982)
19.34 N.H. Fletcher, A. Tarnopolsky: Acoustics of the avian vocal tract, J. Acoust. Soc. Am. **105**, 35–49 (1999)
19.35 F. Goller, O.N. Larsen: A new mechanism of sound generation in songbirds, Proc. Nat. Acad. Sci. U.S.A. **94**, 14787–14791 (1997)
19.36 G.B. Mindlin, T.J. Gardner, F. Goller, R.A. Suthers: Experimental support for a model of birdsong production, Phys. Rev. E **68**, 041908 (2003)
19.37 D.K. Patterson, I.M. Pepperberg: A comparative study of human and parrot phonation: Acoustic and articulatory correlates of vowels, J. Acoust. Soc. Am. **96**, 634–648 (1994)
19.38 G.J.L. Beckers, B.S. Nelson, R.A. Suthers: Vocal-tract filtering by lingual articulation in a parrot, Current Biol. **14**, 1592–1597 (2004)
19.39 O.N. Larsen, F. Goller: Role of syringeal vibrations in bird vocalizations, Proc. R. Soc. London B **266**, 1609–1615 (1999)
19.40 G.J.L. Beckers, R.A. Suthers, C. ten Cate: Pure-tone birdsong by resonant filtering of harmonic overtones, Proc. Nat. Acad. Sci. U.S.A. **100**, 7372–7376 (2003)
19.41 N.H. Fletcher, T. Riede, G.A. Beckers, R.A. Suthers: Vocal tract filtering and the "coo" of doves, J. Acoust. Soc. Am. **116**, 3750–3756 (2004)
19.42 C. Pytte, M.S. Ficken, A. Moiseff: Ultrasonic singing by the blue-throated hummingbird: A comparison between production and perception, J. Comp. Physiol. A **190**, 665–673 (2004)
19.43 G. Sales, D. Pye: *Ultrasonic Communication in Animals* (Chapman Hall, London 1974)

Amazon EX1014 Page 809

19.44 D.J. Hartley, R.A. Suthers: The acoustics of the vocal tract in the horseshoe bat, Rhinolophus hildebrandti, J. Acoust. Soc. Am. **84**, 1201–1213 (1988)
19.45 A.D. Hawkins, A.A. Myrberg: Hearing and sound communication under water. In: *Bioacoustics: A Comparative Approach*, ed. by B. Lewis (Academic, London 1983) pp. 347–405
19.46 N.A.M. Schellart, A.N. Popper: Functional aspects of the evolution of the auditory system in Actinopterygian fish. In: *The Evolutionary Biology of Hearing*, ed. by D.B. Webster, R.R. Fay, A.N. Popper (Springer, Berlin, New York 1992) pp. 295–322
19.47 T.D. Rossing, N.H. Fletcher: *Principles of Vibration and Sound* (Springer, New York 2004), Chapts. 8–10

20. Cetacean Acoustics

The mammalian order cetacea consist of dolphins and whales, animals that are found in all the oceans and seas of the world. A few species even inhabit fresh water lakes and rivers. A list of 80 species of cetaceans in a convenient table is presented by *Ridgway* [20.1]. These mammals vary considerably in size, from the largest living mammal, the large blue whale (*balaenoptera musculus*), to the very small harbor porpoise (*phocoena phocoena*) and Commerson's dolphin (*cephalorhynchus commersonnii*), which are typically slightly over a meter in length.

Cetaceans are subdivided into two suborders, odontoceti and mysticeti. Odontocetes are the toothed whales and dolphins, the largest being the sperm whale (*physeter catodon*), followed by the Baird's beaked whale (*berardius bairdii*) and the killer whale (*orcinus orca*). Within the suborder odontoceti there are four superfamilies: *platanistoidea*, *delphinoidea*, *ziphioidea*, and *physeteridea*. Over half of all cetaceans belong to the superfamily *delphinoidea*, consisting of seven species of medium whales and 35 species of small whales also known as dolphins and porpoises [20.1]. Dolphins generally have a sickle-shaped dorsal fin, conical teeth, and a long rostrum. Porpoises have a more triangular dorsal fin, more spade-shaped teeth, and a much shorter rostrum [20.1].

Mysticetes are toothless, and in the place of teeth they have rigid brush-like whalebone plate material called baleen hanging from their upper jaw. The baleen is used to strain shrimp, krill, micronekton, and zooplankton. All the great whales are mysticetes or baleen whales and all are quite large. The sperm and Baird s beaked whales are the only odontocetes that are larger than the smaller mysticetes such as Minke whales and pygmy right whales. Baleen whales are subdivided into four families, *balaenidae* (right and bowhead whales), *eschrichtiidae* (gray whales), *balaenopteridae* (Minke, sei, Bryde's, blue, fin, and humpback whales), and *neobalaenidae* (pygmy right whale).

Acoustics play a large role in the lives of cetaceans since sound travels underwater better than any other form of energy. Vision underwater is limited to tens of meters under the best conditions and less than a fraction of a meter in turbid and murky waters. Visibility is also limited by the lack of light at great depths during the day and at almost any depth on a moonless night. Sounds are used by marine mammals for myriad reasons such as group cohesion, group coordination, communications, mate selection, navigation and locating food. Sound is also used over different spatial scales from tens of km for some species and tens of meters for other species, emphasizing the fact that different species utilize sound in different ways. All odontocetes seem to have the capability to echolocate, while mysticetes do not echolocate except in a very broad sense, such as listening to their sound bouncing off the bottom, sea mounts, underwater canyon walls, and other large objects.

The general rule of thumb is that larger animals tend to emit lower-frequency sounds and the frequency range utilized by a specific species may be dictated more from anatomical constraints than any other factors. If resonance is involved with sound production, then anatomical dimensions become critical, that is, larger volumes resonate at lower frequencies than smaller volumes. The use of a particular frequency band will also have implications as to the distance other animals, including conspecifics, will be able to hear the sounds. Acoustic energy is lost in the propagation process by geometric spreading and absorption. Absorption losses are frequency dependent, increasing with frequency. Therefore, the sounds of baleen whales such as the blue whale that emit sounds with fundamental frequencies as low as 15 Hz can propagate to much longer distances than the whistles of dolphins that contain frequencies between 5 and 25 kHz.

20.1 **Hearing in Cetaceans** 806
20.1.1 Hearing Sensitivity of Odontocetes 807

Amazon EX1014 Page 811

20.1.2 Directional Hearing in Dolphins .. 808
20.1.3 Hearing by Mysticetes 812
20.2 **Echolocation Signals** 813
20.2.1 Echolocation Signals of Dolphins that also Whistle 813
20.2.2 Echolocation Signals of Smaller Odontocetes that Do not Whistle . 817
20.2.3 Transmission Beam Pattern 819
20.3 **Odontocete Acoustic Communication** 821
20.3.1 Social Acoustic Signals 821
20.3.2 Signal Design Characteristics 823
20.4 **Acoustic Signals of Mysticetes** 827
20.4.1 Songs of Mysticete Whales 827
20.5 **Discussion** 830
References 831

20.1 Hearing in Cetaceans

One of the obvious adaptations for life in the sea is the absence of a pinna in cetaceans. The pinna probably disappeared through a natural selection process because it would obstruct the flow of water of a swimming animal and therefore be a source of noise. In the place of a pinna, there is a pin-hole on the surface of a dolphin's head which leads to a ligament inside the head, essentially rendering the pinna nonfunctional in conducting sounds to the middle ear. So, how does sounds enter into the heads of cetacean? Several electrophysiological experiments have been performed in which a hydrophone is held at different locations on an animal's head and the electrophysiological thresholds are determined as a function of the position of the hydrophone [20.2–4]. All three studies revealed greatest sensitivity on the dolphin's lower jaw.

The experimental configuration of *Møhl* et al. [20.2] is shown in Fig. 20.1a, which shows a *suction-cup hydrophone* and attached a bottlenose dolphin's lower jaw and the surface contact electrodes embedded in suction cups used to measure the auditory brainstem potential signals for different levels of sound intensity. The important differences in the experiment of *Møhl* et al. [20.2] are that the subject was trained and the measurements were done in air so that the sounds were limited to the immediate area where they were applied. The results of the experiment are shown in Fig. 20.1b. The location of maximum sensitivity to sound is slightly forward of the pan-bone area of the lower jaw, a location where *Norris* [20.5] hypothesized that sounds enter the head of a dolphin. Numerical simulation work by *Aroyan* [20.6] suggest that sounds enter the dolphin's head forward of the pan bone, but a good portion of the sound propagates below the skin to the pan bone and enters the head through the pan bone. The fact that sounds probably propagate through the lower jaw of dolphins and other

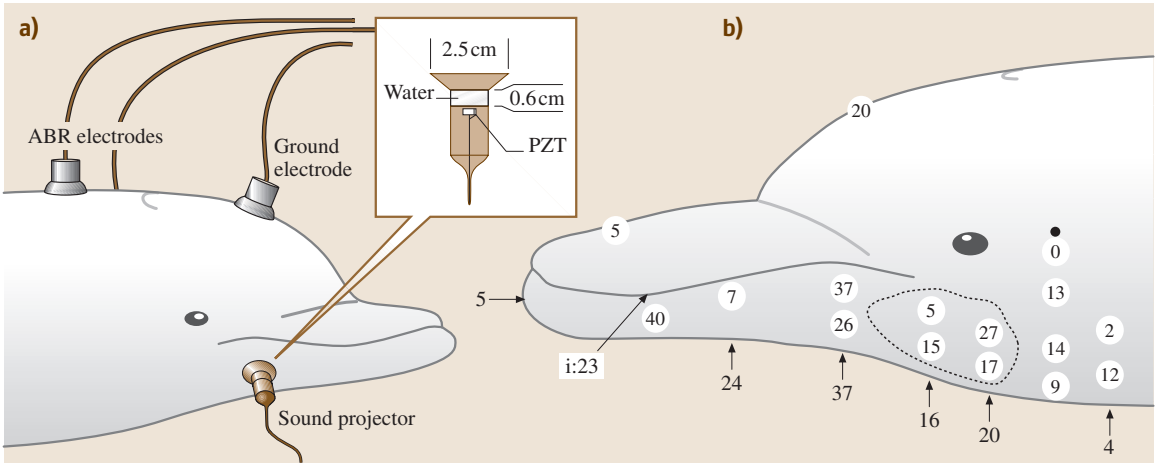


Fig. 20.1 **(a)** The experimental geometry used by *Møhl* et al. [20.2], **(b)** results of the auditory brainstem response (ABR) threshold measurements. The numerical values represent the amount of attenuation of the sound needed to obtain an ABR threshold. Therefore, the higher the number the more sensitive the location

Amazon EX1014 Page 812

odontocetes does not necessarily mean that the same or a similar propagation process is occurring with baleen whales.

20.1.1 Hearing Sensitivity of Odontocetes

Almost all our knowledge of hearing in cetaceans comes from studies performed with small odontocetes. The most studied species is the Atlantic bottlenose dolphin (*tursiops truncatus*). Despite the amount of research performed with the bottlenose dolphin, our understanding of auditory processes in these animals lags considerably behind that for humans and other terrestrial mammals. There are still many large gaps in our knowledge of various auditory processes occurring within the most studied odontocetes. The first audiogram for a cetacean was measured by *Johnson* [20.9] for a *tursiops truncatus*. Since then, audiograms have been determined for the harbor porpoise (*phocoena phocoena*) by *Andersen* [20.10] and *Kastelein* et al. [20.11], the killer whale (*orcinus orca*) by *Hall* and *Johnson* [20.12] and *Szymanski* et al. [20.13], the beluga whale (*delphinapterus leucas*) by *White* et al. [20.14], the Pacific bottlenose dolphin (*tursiops gilli*) by *Ljungblad* et al. [20.15], the false killer whale (*pseudorca crassidens*) by *Thomas* et al. [20.16], the Chinese river dolphin (*lipotes vexillifer*) by *Wang* et al. [20.17], Risso's dolphins (*grampus griseus*) by *Nachtigall* et al. [20.18], the tucuxi (*sotalia fluviatilis*) by *Sauerland* and *Dehnhardt* [20.19], and the striped dolphin (*stenella coeruleoalba*) by *Kastelein* et al. [20.20]. The audiograms of these odontocetes are shown in Fig. 20.2. It is relatively striking to see how similar the audiograms are between species considering the vastly different habitats and areas of the world where some of these animals are found and the large differences in body size. All the audiograms suggest high-frequency hearing capabilities, with the smallest animal, *phocoena phocoena* having the highest hearing limit close to 180 kHz. However, the *orcinus orca*, which is over 95 times heavier and about six times longer can hear up to about 105 kHz. The actual threshold values shown in Fig. 20.2 should not be compared between species because the different methods of determining the threshold can lead to different results and because of the difficulties of obtaining good sound pressure level measurements in a reverberant environment. For example, *Kastelein* et al. [20.11] used a narrow-band frequency-modulated (FM) signal to avoid multi-path problems and the FM signals may provide additional cues not present in a pure-tone signal. Nevertheless, the audiograms shown in Fig. 20.2 suggest that all the animals had similar thresholds of 10–15 dB.

A summary of some important properties of the different audiograms depicted in Fig. 20.2 is given in Table 20.1. In the table, the frequency of best hearing is arbitrary defined as the frequency region in which the auditory sensitivity is within 10 dB of the maximum sensitivity depicted in each audiogram of Fig. 20.2. With the exception of the *Orcinus* and the *Lipotes*, the maximum sensitivity of the rest of the species represented

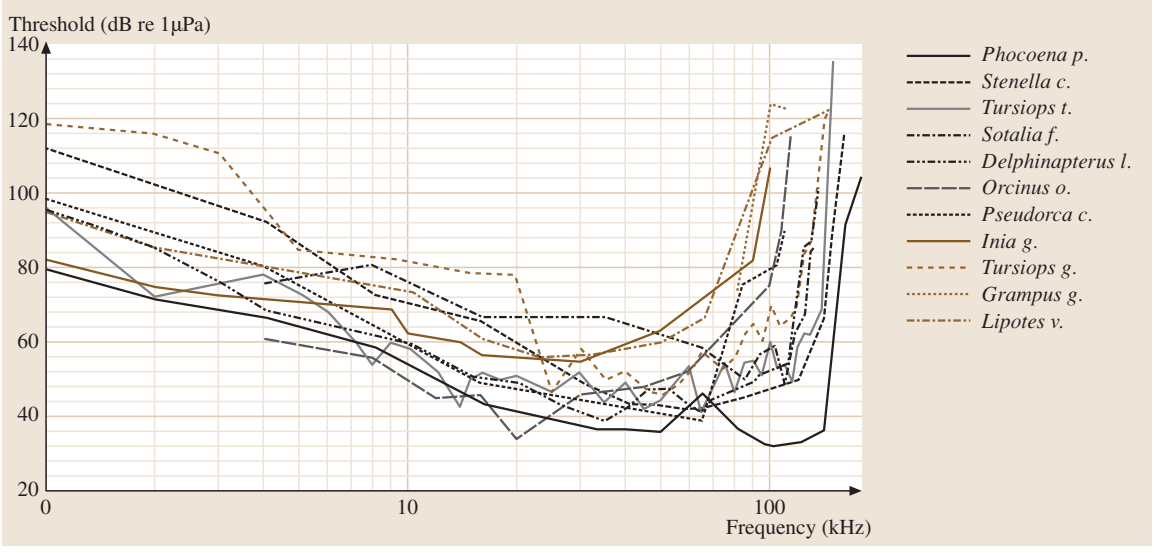


Fig. 20.2 Audiogram for different odontocetes species (after [20.7, 8])

Table 20.1 Some important properties of the audiograms plotted in Fig. 20.2

Species	Maximum sensitivity (dB re 1 : Pa)	Frequency of best hearing (kHZ)	Upper frequency limit (kHz)
phocoena phocoena	32	18 – 130	180
stenella coeruleoalba	42	30 – 125	160
tursiops truncatus	42	15 – 110	150
tursiops gilli	47	30 – 80	135
sotalia fluviatilis	50	35 – 50	135
delphinapterus leucas	40	11 – 105	120
pseudorca crassidens	39	17 – 74	115
orcinus orca	34	15 – 40	110
inia geoffrensis	51	12 – 64	100
lipotes vexillifer	55	15 – 60	100
grampus griseus	–	–	100

in Fig. 20.2 and Table 20.1 are very similar, and within experimental uncertainties, especially for audiograms obtained with the staircase procedure using relatively large step sizes of 5 dB or greater. At the frequency of best hearing, the threshold for *Orcinus* and *Phocoena* are much lower than for the other animals. It is not clear whether this keen sensitivity is a reflection of a real difference or a result of some experimental artifact. The data of Table 20.1 also indicate that *Phocoena phocoena*, *Stenella coeruleoalba* and *Tursiops truncatus* seem to have the widest auditory bandwidth.

20.1.2 Directional Hearing in Dolphins

Sound Localization

The ability to localize or determine the position of a sound source is important in order to navigate, detect prey and avoid predators and avoid hazards producing an acoustic signatures. The capability to localize sounds has been studied extensively in humans and in many vertebrates (see [20.21]). Lord Rayleigh in 1907 proposed that humans localized in the horizontal plane by using interaural time differences for low-frequency sounds and interaural intensity differences for high-frequency sounds. The sound localization acuity of a subject is normally defined in terms of a minimum audible angle (MAA), defined as the angle subtended at the subject by two sound sources, one being at a reference azimuth, at which the subject can just discriminate the sound sources as two discrete sources [20.22]. If the sound sources are separated symmetrically about a midline, the MAA is one half the threshold angular separation between the sound sources. If one of the sound sources is located at the midline, then the MAA is the same as the threshold angular separation between the two sources.

Renaud and *Popper* [20.23] examined the sound localization capabilities of a *tursiops truncatus* by measuring the MAA in both the horizontal and vertical planes. During a test trial the dolphin was required to station on a bite plate facing two transducers positioned at equal angles from an imaginary line running through the center of the bite plate. An acoustic signal was then transmitted from one of the transducers and the dolphin was required to swim and touch the paddle on the same side as the emitting transducer. The angle between the transducers was varied in a modified staircase fashion. If the dolphin was correct for two consecutive trials, the transducers were moved an incremental distance closer together, decreasing the angle between the transducer by 0.5°. After each incorrect trial, the transducers were moved an incremental distance apart, increasing the angle between the transducers by 0.5°. This modified staircase procedure allowed threshold determination at the 70% level. The threshold angle was determined by averaging a minimum of 16 reversals. A randomized schedule for sound presentation through the right or left transducer was used.

The localization threshold determined in the horizontal and vertical planes as a function of frequency for narrow-band pure-tone signals is shown in Fig. 20.3. The MAA had a U-shaped pattern, with a large value of 3.6° at 6 kHz, decreasing to a minimum of 2.1° at 20 kHz and then slowly increasing in an irregular fashion to 3.8° at 100 kHz. MAAs for humans vary between 1° and 3.7°, with the minimum at a frequency of approximately 700 Hz [20.24]. The region where the MAA decreased to a minimum in Figs. 20.19 and 20.20 (about 20 kHz) may be close to the frequency at which the dolphin switches from using interaural time difference cues to interaural intensity difference cues. The MAAs for the

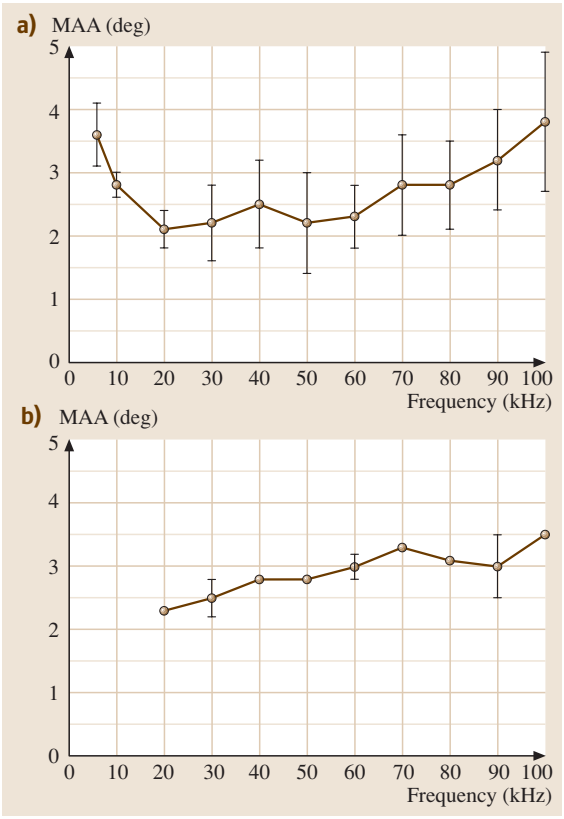


Fig. 20.3 **(a)** Localization threshold determined in the horizontal plane as a function of frequency. The mean ± one standard deviation is shown for seven determinations per frequency. The animal faced directly ahead at 0° azimuth. **(b)** Localization threshold determined in the vertical plane as a function of frequency. Standard deviation are indicated for 30, 60 and 90 kHz vertical data (seven sessions each). The dolphin's azimuth was 0° (after *Renauld* and *Popper*, [20.27]) Figs. 20.19 and 20.20

bottlenose dolphin were smaller than the MAAs (in the horizontal plane) of 3.5° at 3.5 kHz and 6° at 6 kHz for a harbor porpoise measured by *Dudok van Heel* [20.25], and 3° at 2 kHz, measured by *Andersen* [20.26] also for a harbor porpoise.

In order to measure the MAA in the vertical plane, the dolphin was trained to turn to its side (rotate its body 90° along its longitudinal axis) and to bite on a sheet of plexiglass which was used as the stationing device. The MAA in the vertical plane varied between 2.3° at 20 kHz to 3.5° at 100 kHz. These results indicate that the animal could localize in the vertical plane nearly as well as in the horizontal plane. These results were not expected since binaural affects, whether interaural time or intensity, should not be present in the vertical plane. However, the dolphin's ability to localize sounds in the vertical plane may be explained in part by the asymmetry in the receive beam patterns in the vertical plane discussed in the next section.

Renaud and *Popper* [20.23] also determined the MAA for a broadband transient signal or click signal, having a peak frequency of 64 kHz and presented to the dolphin at a repetition rate of 333 Hz. The MAA in the horizontal plane was found to be approximately 0.9° and 0.7° in the vertical plane. It is not surprising that a broadband signal should result in a lower MAA than a pure-tone signal of the same frequency as the peak frequency of the click. The short onset time and the broad frequency spectrum of a click signal should provide additional cues for localization.

Receiving Beam Patterns

Having narrow transmission and reception beams allows the dolphin to localize objects in a three-dimensional volume, spatially separate objects within a multi-object field, resolve features of extended or elongated objects, and to minimize the amount of interference received. The amount of ambient noise from an isotropic noise field and the amount of reverberation interference received is directly proportional to the width of the receiving beam. The effects of discrete or partially extended interfering noise or reverberant sources can be minimized by simply directing the beams away from the sources.

The receiving beam pattern of a dolphin was determined by measuring the masked hearing threshold as a function of the azimuth about the animal's head. The relative masked hearing threshold as a function of azimuth is equivalent to the received beam pattern since the receiving beam pattern is the spatial pattern of hearing sensitivity. *Au* and *Moore* [20.28] measured the dolphin's masked hearing threshold as the position of either the noise or signal sources varied in their angular position about the animal's head. The dolphin was required to voluntarily assume a stationary position on a bite plate constructed out of a polystyrene plastic material. The noise and signal transducers were positioned along an arc with the center of the arc located approximately at the pan bone of the animal's lower jaw. In order to measure the dolphin's receiving beam in the vertical plane, the animal was trained to turn onto its side before biting the specially designed vertical bite-plate stationing device. For the measurement in the vertical plane, the position of the signal transducer was fixed directly in

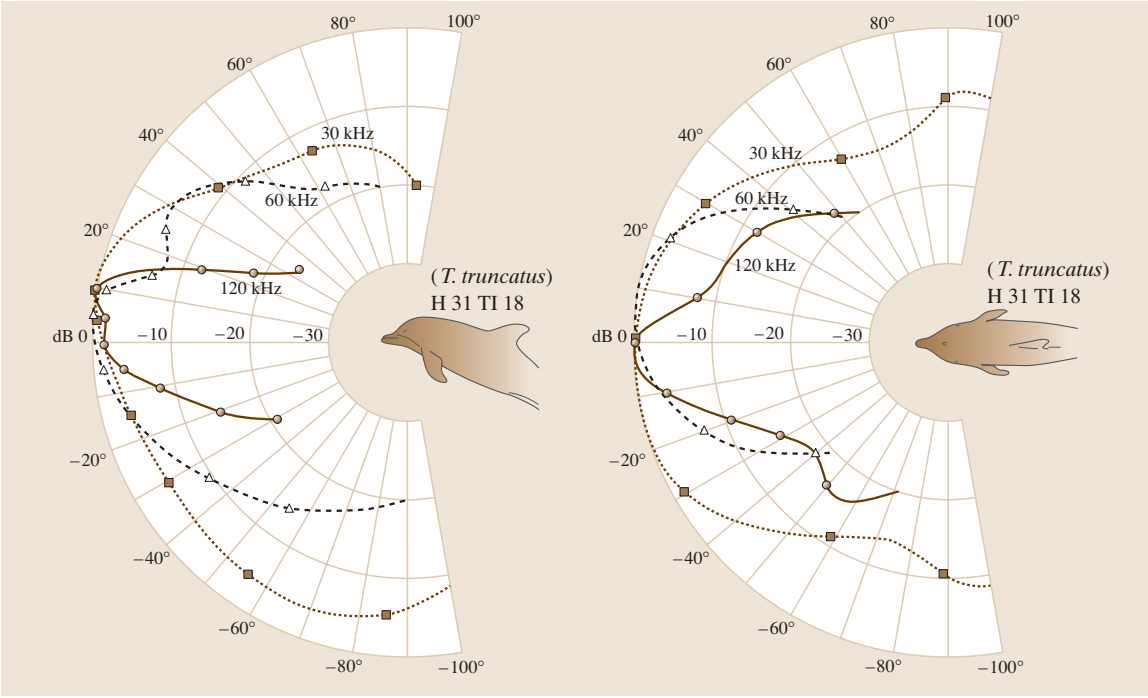


Fig. 20.4 Receive beam patterns in the vertical and horizontal planes for frequencies of 30, 60 and 120 kHz. The relative masked thresholds as a function of the elevation angle of the asking noise source are plotted for each signal frequency (after *Au* and *Moore*, [20.28])

line with the bite plate and its acoustic output was held constant. Masked thresholds were measured for different angular position of the noise transducer along the arc. The level of the noise was varied in order to obtain the masked threshold. A threshold estimate was considered complete when at least 20 reversals (10 per session) at a test angle had been obtained over at least two consecutive sessions, and if the average reversal values of the two sessions were within 3 dB. After a threshold estimate was achieved, the noise transducer was moved to a new azimuth over a set of randomly predesignated azimuths. As the azimuth about the dolphin's head increased, the hearing sensitivity of the dolphin tended to decrease, requiring higher levels of masking noise from a transducer located at that azimuth to mask the signal from a source located directly ahead of the animal.

The receiving beam patterns in both the vertical and horizontal plane are plotted for signal frequencies of 30, 60 and 120 kHz in Fig. 20.4. The radial axis of Fig. 20.4 represents the difference in dB between the noise level needed to mask the test signal at any azimuth and the minimum noise level needed to mask the test signal at the azimuth corresponding to the major axis of the vertical beam. The shape of the beams in Fig. 20.4 indicates that the patterns were dependent on frequency, becoming narrower, or more directional as the frequency increased. The beam of a planar hydrophone also becomes narrower as frequency increases. The 3 dB beam widths were approximately 30.4°, 22.7°, and 17.0° for frequencies of 30, 60, and 120 kHz, respectively. There was also an asymmetry between the portion of the beam above and below the dolphin's head. The shape of the beams dropped off more rapidly as the angle above the animal's head increased than for angles below the animal's head, indicating a more rapid decrease in the animal's hearing sensitivity for angles above the head than for angles below the head. If the dolphin receives sounds through the lower jaw, the more rapid reduction in hearing sensitivity for angles above the head may have been caused by shadowing of the received sound by the upper portion of the head structure including air in the nasal sacs [20.29]. There is a slight peculiarity in the 60 kHz beam which shows almost the same masked threshold values for 15° and 25° elevation angles.

The radial line passing through the angle of maximum sensitivity is commonly referred to as the major

axis of the beam. The major axis of the vertical beams is elevated between 5° and 10° above the reference axis. The 30 and 120 kHz results show the major axis at 10° while the 60 kHz results showed the major axis at 5°. It will be shown in a later section that the major axis of the received beam in the vertical plane is elevated at approximately the same angle as the major axis of the transmitted beam in the vertical plane.

In the horizontal beam pattern measurement, two noise sources were fixed at azimuth angles of ±20°. The level of the noise sources was also fixed. The position of the signal transducer was varied from session to session. Masked thresholds were determined as a function of the azimuth of the signal transducer, by varying the signal level of the signal transducer in a staircase fashion. A threshold estimate was considered completed when at least 20 reversals at a test angle had been obtained over at least two consecutive sessions, if the average reversal values were within 3 dB of each other. After a threshold estimate was determined, the signal transducer was moved to a new azimuth over a set of randomly predesignated azimuths. Two noise sources were used in order to discourage the dolphin from internally steering its beam in the horizontal plane. If the animal could steer its beam, it would receive more noise from one of the two hydrophones, and therefore not experience any improvement in the signal-to-noise ratio. The masking noise from the two sources was uncorrelated but equal in amplitude. The radial axis represents the difference in dB between the signal level at the masked threshold for the various azimuths and the signal level of the masked threshold for 0° azimuth (along the major axis of the horizontal beam). The horizontal receiving beams were directed forward with the major axis being parallel to the longitudinal axis of the dolphin. The beams were nearly symmetrical about the major axis. Any asymmetry was within the margin of experimental error involved in estimating the relative thresholds. The horizontal beam patterns exhibited a similar frequency dependence as the vertical beam patterns, becoming narrower or more directional as the frequency increased. The 3 dB beam widths were 59.1°, 32.0°, and 13.7° for frequencies of 30, 60, and 120 kHz, respectively.

Zaytseva et al. [20.30] measured the horizontal beam pattern of a dolphin by measuring the masked hearing threshold as a function of azimuth. Their beam width of 8.2° for a frequency of 80 kHz was much narrower than the 13.7° for a frequency of 120 kHz. The difference in beam width is even larger if the results of *Au* and *Moore* [20.28] are linearly interpolated to 80 kHz. We calculated an interpolated beam width of 25.9° at 80 kHz, which was considerably greater than the 8.2° obtained by *Zaytseva* et al. [20.30]. The difference in beam width measured by Zaytseva et al. and Au and Moore may be attributed to the use of only one noise source by Zaytseva et al. compared to the two noise sources used by Au and Moore. With a single masking noise source in the horizontal plane, there is the possibility of the animal performing a spatial filtering operation by internally steering the axis of its beam in order to maximize the signal-to-noise ratio. Another possibility is that Zaytseva et al. did not use a fixed stationing device. Rather, the dolphin approached the signal hydrophone from a start line, always oriented in the direction of the signal hydrophone. The animal responded to the presence or absence of a signal by either swimming or not swimming to the hydrophone. In such a procedure, it is impossible to control the orientation of the animal's head with respect to the noise masker so that the dolphin could move its head to minimize the effects of the noise.

Directivity Index

The directivity index is a measure of the sharpness of the beam or major lobe of either a receiving or transmitting beam pattern. For a spherical coordinate system, the directivity index of a transducer is given by the equation [20.31]

$$\mathrm{DI} = 10\log\frac{4\pi}{\int_0^{2\pi}\int_{-\pi/2}^{\pi/2}\left(\frac{p(\theta,\phi)}{p_0}\right)^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi}\,. \qquad (20.1)$$

Although the expression for directivity index is relatively simple, using it to obtain numerical values can be quite involved unless transducers of relatively simple shapes (cylinders, lines and circular apertures) with symmetry about one axis is involved. Otherwise, the beam pattern needs to be measured as a function of both θ, and ϕ. This can be done by choosing various discrete values of θ and measuring the beam pattern as a function of ϕ, a tedious process. Equation (20.1) can then be evaluated by numerically evaluating the double integral with a digital computer. The directivity indices associated with the dolphin's beam patterns in Fig. 20.4 were estimated by *Au* and *Moore* [20.28] using (20.1) and a two-dimensional Simpson's 1/3-rule algorithm [20.32]. The results of the numerical evaluation of are plotted as a function of frequency in Fig. 20.5. DIs of 10.4, 15.3 and 20.6 dB were obtained for frequencies of 30, 60,

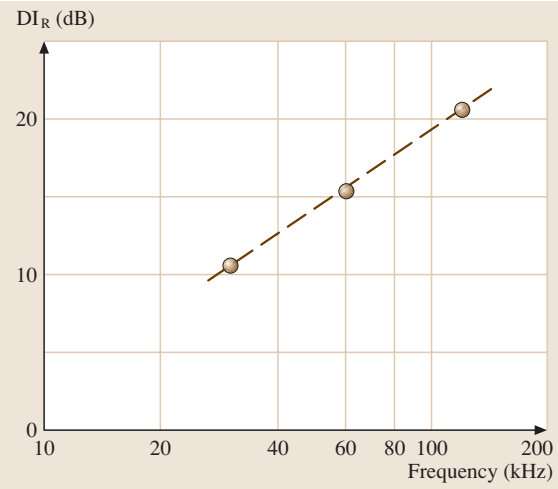


Fig. 20.5 Receiving directivity index as function of frequency for a *tursiops truncatus*

and 120 kHz, respectively. A linear curve fitted to the computed DIs in a least-square-error manner is also shown in the Fig. 20.5. The equation of the line is

$$\mathrm{DI(dB)} = 16.9 \log f(\mathrm{kHz}) - 14.5 \,. \qquad (20.2)$$

The results of Fig. 20.5 indicate that the dolphin's receive directivity index increased with frequency in a manner similar to that of a linear transducer (Bobber, 1970). The expression in (20.2) is only valid for frequencies at which DI(dB) is greater or equal to 0.

Although the directivity index expressed by (20.1) is for a *tursiops*, it can also be used to estimate the directivity index of other dolphins by applying an appropriate correction factor, so that (20.2) can be rewritten as

$$\mathrm{DI(dB)} = 16.9 \log f(\mathrm{kHz}) - 14.5 + \mathrm{CF(dB)} \,, \qquad (20.3)$$

where CF(dB) is a correction factor taking into account different head sizes. The directivity index of a planar circular plate is proportional to its diameter so if we let d_T be the diameter of the head of a *tursiops* at about the location of the blowhole and d_D be the diameter of the head of a different species of dolphin, then the correction factor can be expressed as

$$\mathrm{CF(dB)} = 20 \log(d_D/d_T) \,. \qquad (20.4)$$

The correction factor will be positive for a dolphin with a larger head and negative for a dolphin with a smaller head than *tursiops*.

20.1.3 Hearing by Mysticetes

Our knowledge of the hearing characteristics of baleen whales is extremely limited. We do not know how they receive sounds, the frequency range of hearing, and the sensitivity of hearing at any frequency. Much of our understanding of hearing in baleen whales comes from anatomical studies of the ears of different species. Baleen whales have occluded external auditory canals that are filled with a homogeneous wax [20.33]. The lower jaws of mysticetes are designed for sieving or gulp feeding and have no evident connection to the temporal bones [20.33] making it very difficult to understand how sounds enter into the ears of these whales.

Various types of whales have been observed by many investigators to react strongly and drastically change their behavior in the presence of boats and low-flying aircraft, however, the sound pressure levels at the whales' locations are often difficult to define and measure. In some situations, the sound levels of the aversive sound could be estimated and a data point obtained. Bowhead whales were observed fleeing from a 13 m diesel-powered boat having a noise level at the location of the whales of about 84 dB re 1 μPa in the dominant 1/3-octave band, or about 6 dB above the ambient noise in that band [20.34]. Playback experiments with bowhead whales indicated that they took evasive action when the noise was about 110 dB re 1 μPa or 30 dB above the ambient noise in the same 1/3-octave band [20.35]. Playback experiments with migrating gray whales indicated that about 10% of the whales made avoidance behavior when the noise was about 110 dB in a 1/3 octave band, 50% at about 117 dB and 90% at about 122 dB or greater. These playback signals consisted of anthropogenic noise associated with the oil and gas industry.

Frankel [20.36] played back natural humpback whale sounds and a synthetic sound to humpback whales wintering in the waters of the Hawaiian islands. Twenty seven of the 1433 trials produced rapid approach response. Most of the responses were to the feeding call. Feeding call and social sounds produced changes in both separation and whale speed, indicating that these sounds can alter a whale's behavior. The humpback whales responded to sounds as low as 102–105 dB but the strongest responses occurred when the sounds were 111 to 114 dB re 1 μPa.

All of the playback experiments suggest that sounds must be between 85 and 120 dB before whales will react to them. These levels are very high compared to those that dolphins can hear and may suggest that it is very

difficult to relate reaction to hearing sensitivity. Whales may not be reacting strongly unless the sounds are much higher than their hearing threshold.

Determining the hearing sensitivity or audiogram of baleen whales represents an extremely difficult challenge and will probably require the use of some sort of electrophysiological technique as suggested by *Ridgway* et al. [20.37]. Perhaps a technique measuring auditory-evoked potentials with beached whales may provide a way to estimate hearing sensitivity. Even with evoked potential measurements, there are many issues that have to be considered. For example, if an airborne source is used, the results cannot be translated directly to the underwater situation. If a sound source is placed on a whale's head, relating that to a whale receiving a plane wave will also not be simple. Measurement of evoked potentials themselves may not be simple because of the amount of flesh, muscles and blubber that the brain waves would have to travel through in order to reach the measurement electrodes. *Ridgway* and *Carder* [20.38] reported on some attempts to use the evoked potential method with a young gray whale held at Sea World in California. They also showed that the pygmy whale auditory system was most sensitive to very high frequencies (120–130 kHz) in the same range as their narrow-band echolocation pulses. A young sperm whale was most sensitive in the 10–20 kHz region of the spectrum [20.38].

20.2 Echolocation Signals

Echolocation is the process in which an organism projects acoustic signals and obtains a sense of its surroundings from the echoes it receives. In a general sense, any animal with a capability to hear sounds can echolocate by emitting sounds and listening to the echoes. A person in an empty room can gain an idea of the size and shape of the room by emitting sounds and listening to the echoes from the different walls. However, in this chapter, echolocation is used in a more specific sense in which an animal has a very specialized capability to determine the presence of objects considerably smaller than itself, discriminate between various objects, recognize specific objects and localize objects in three-dimensional space (determine range and azimuth). Dolphins and bats have this specialized capability of echolocation.

The echolocation system of a dolphin can be broken down into three subsystems: the transmission, reception and signal processing/decision subsystems. The reception system has to do with hearing and localization. The transmission system consist of the sound production mechanism, acoustic propagation from within the head of the dolphin to into the water, and the characteristics of the signals traveling in the surrounding environment. The third subsystem has to do with processing of auditory information by the peripheral and central nervous system. The capability of a dolphin to detect objects in noise and in clutter, and to discriminate between various objects depends to a large extent on the information-carrying capabilities of the emitted signals.

Dolphins most likely produce sounds within their nasal system and the signals are projected out through the melon. Although there has been a long-standing controversy on whether sounds are produced in the larynx or in the nasal system of odontocetes, almost all experimental data with dolphins indicate that sounds are produced in the nasal system [20.39]. For example, *Ridgway* and *Carder* [20.40] used catheters accepted into the nasal cavity by trained belugas. The belugas preformed an echolocation task in open water, detecting targets and reported the presence of targets by whistling. Air pressure within the nasal cavity was shown to be essential for echolocation and for whistling [20.40].

The melon immediately in front of the nasal plug may play a role in channeling sounds into the water, a notion first introduced by *Wood* [20.41]. *Norris* and *Harvey* [20.42] found a low-velocity core extending from just below the anterior surface towards the right nasal plug, and a graded outer shell of high-velocity tissue. Such a velocity gradient could channel signals originating in the nasal region in both the vertical and horizontal planes. Using both a two-dimensional [20.43] and a three-dimensional model [20.6] to study sound propagation in a dolphin's head, Aroyan has shown that echolocation signals most likely are generated in the nasal system and are channeled into the water by the melon. *Cranford* [20.44] has also collected evidence from nasal endoscopy of trained echolocating dolphins that suggest that echolocation signals are most likely produced in the nasal system at the location of the monkey-lips, dorsal bursae complex just beneath the blow hole.

Amazon EX1014 Page 819

20.2.1 Echolocation Signals of Dolphins that also Whistle

Most dolphin species are able to produce whistle signals. Among some of the species in this category in which echolocation signals have been measured include the bottlenose dolphin (*tursiops sp*), beluga whale (*delphinapterus leucas*), killer whale (*orcinus orca*), false killer whale (*pseudorca crassidens*), Pacific white-sided dolphin (*lagenorhynchus obliquidens*), Amazon river dolphin (*inia geoffrensis*), Risso's dolphin (*grampus griseus*), tucuxi (*sotalia fluviatilis*), Atlantic spotted dolphin (*stenella frontalis*), Pacific spotted dolphin (*Stenella attenuata*), spinner dolphin (*Stenella longirostris*), pilot whale (*globicephala sp*), rough-toothed dolphin (*steno bredanesis*), Chinese river dolphin (*lipotes vexillifer*) and sperm whales (*physeter catodon*). However, most of the available data have been obtained for three species: the bottlenose dolphin, the beluga whale and the false killer whale.

Prior to 1973, most echolocation signals of *tursiops* were measured in relatively small tanks and the results provided a completely different picture of what we currently understand. It was not until the study of *Au* et al. [20.45] that certain features of biosonar signals used by *tursiops* and other dolphins in open waters were discovered. We discovered that the signals had peak frequencies between 120 and 130 kHz, over an octave higher than previously reported peak frequencies between 30 and 60 kHz [20.46]. We also measured an average peak-to-peak click source level on the order of 220 dB re 1 μPa at 1 m, which represents a level 30 to 50 dB higher than previously measured for *tursiops*. Examples of typical echolocation signals emitted by *tursiops truncatus* are shown in Fig. 20.6 for two situations. The top signal is typical of signals used in the open waters of Kaneohe Bay, Oahu, Hawaii, and the second signal represents typical signals for a *tursiops* in a tank. Signals measured in Kaneohe Bay regularly have duration between 40 and 70 μs, 4 to 10 positive excursion, peak frequencies between 110 and 130 kHz and peak-to-peak source levels between 210 and 228 dB re 1 μPa. The signals in Fig. 20.6 are not drawn to scale; if they were, the tank signal would resemble a flat line.

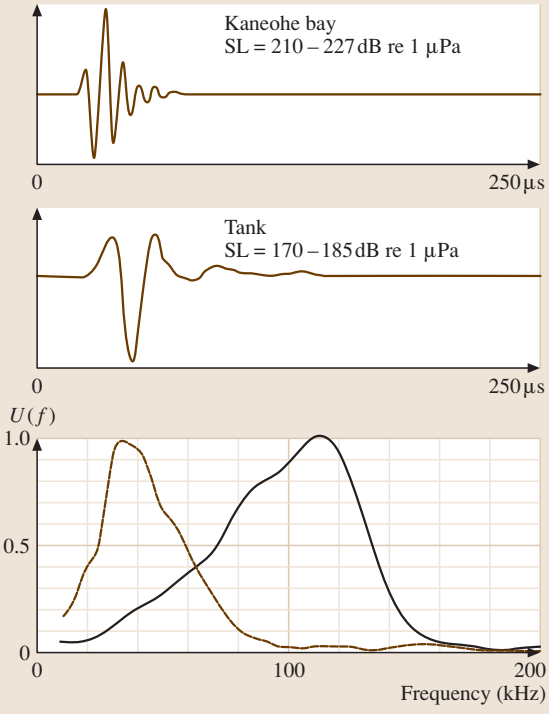


Fig. 20.6 Example of echolocation signals used by *tursiops truncatus* in Kaneohe Bay (after *Au* [20.45]), and in a tank

Au et al. [20.47] postulated that high-frequency echolocation signals were a byproduct of the animals producing high-intensity clicks to overcome snapping shrimp noise. In other words, dolphins can only emit high-level clicks (greater than 210 dB) if they use high frequencies. The effects of a noisy environment on the echolocation signals used by a beluga or white whale (*Delphinapterus leucas*) was vividly demonstrated by *Au* et al. [20.47]. The echolocation signal of a beluga was measured in San Diego Bay, California before the whale was moved to Kaneohe Bay. The ambient noise in Kaneohe Bay is between 15 to 20 dB greater than in San Diego Bay. The whale emitted echolocation signals with peak frequencies between 40 and 60 kHz and with a maximum averaged peak-to-peak source level of 202 dB re 1 μPa in San Diego Bay. In Kaneohe Bay, the whale shifted the peak frequency of its signals over an octave higher to 100 and 120 kHz. The source level also increased to over 210 dB re 1 μPa [20.47]. Examples of typical echolocation signals used by the whale in San Diego and in Kaneohe Bay are shown in Fig. 20.7a. Here, the signals are drawn to scale with respect to each other. Echolocation signals used by belugas in tanks also resemble the low-frequency signals shown in Fig. 20.2a [20.48, 49]. *Turl* et al. [20.50] measured the sonar signals of a beluga in a target-in-clutter detection task in San Diego Bay and found the animal used high-frequency (peak frequency above 100 kHz) and high-intensity (greater than 210 dB re 1 μPa) signals. Therefore low-amplitude clicks of the beluga had

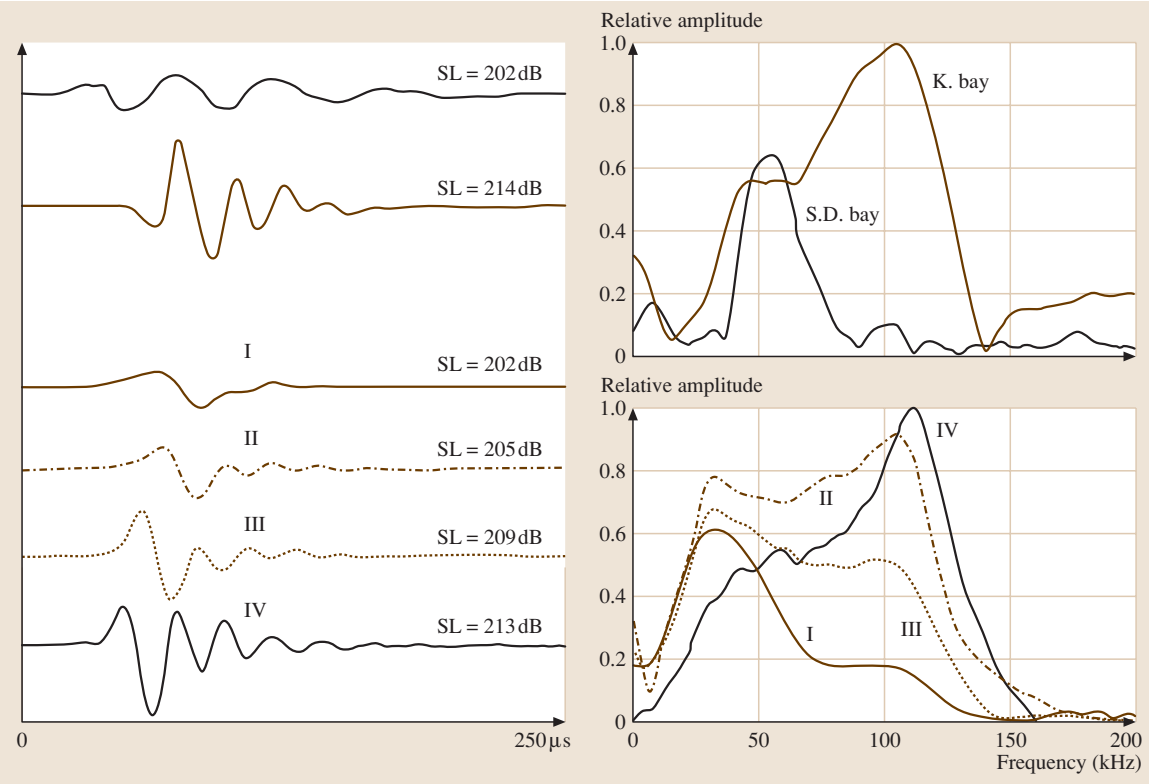


Fig. 20.7 **(a)** Example of beluga echolocation signals measured in San Diego Bay and in Kaneohe Bay (after *Au* et al. [20.47]); **(b)** Examples of *pseudorca* echolocation signals, Sl is the averaged peak-to-peak source level (after *Au* et al. [20.52])

low peak frequencies and the high-amplitude clicks had high peak frequencies. The data of *Moore* and *Pawloski* [20.51] for *Tursiops* also seem to support the notion that the shape of the signal in the frequency domain is related to the intensity of the signal.

Recent results with a false killer whale showed a clear relationship between the frequency content of echolocation signals and source level [20.52]. The *Pseudorca* emitted four basic types of signals, which are shown in Fig. 20.7b. The four signal types have spectra that are bimodal (having two peaks); the spectra in Fig. 20.2 are also bimodal. The type I signals were defined as those with the low-frequency peak (< 70 kHz) being the primary peak and the high-frequency peak being the secondary peak with its amplitude at least 3 dB below that of the primary peak. Type II signals were defined as those with a low-frequency primary peak and a high-frequency secondary peak having an amplitude within 3 dB of the primary peak. Type III signals were those with a high-frequency primary peak (>70 kHz), and a low-frequency secondary peak having an amplitude within 3 dB of the primary peak. Finally, type IV signals were those with a high-frequency primary peak having an amplitude that was at least 3 dB higher than that of the secondary low-frequency peak.

The data of *Thomas* et al. [20.53, 54] also indicated a similar relationship between intensity and the spectrum of the signal. The echolocation signals of a *pseudorca* measured in a tank had peak frequencies between 20 and 60 kHz and source levels of approximately 180 dB re 1 µPa [20.53]. Most of the sonar signals used by another *pseudorca* performing a detection task in the open waters of Kaneohe Bay had peak frequencies between 100 and 110 kHz and source levels between 220 and 225 dB re 1 µPa [20.54].

A bimodal spectrum is best described by its center frequency, which is defined as the centroid of the spectrum, and is the frequency which divides the spectrum into two parts of equal energy. From Fig. 20.7, we can see that, as the source level of the signal in-

Amazon EX1014 Page 821

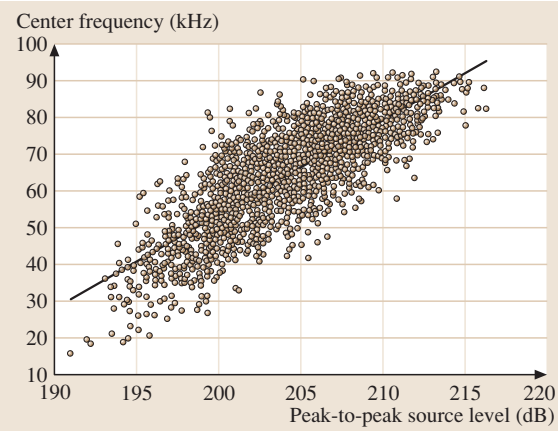


Fig. 20.8 Center frequency of echolocation signals emitted by a *pseudorca* as a function of the peak-to-peak source level (after *Au* et al. [20.52])

creased, the frequency components at higher frequencies also increased in amplitude, suggesting a relationship between source level and center frequency. This relationship can be examined by considering the scatter diagram of Fig. 20.8 showing center frequency plotted against source level. The solid line in the figure is a linear-regression curve fit of the data and has a correlation coefficient of 0.80.

The bimodal property of the echolocation signals of Fig. 20.7 seems to suggest that the response of the sound generator may be determined by the intensity of the driving force that eventually causes an echolocation signal to be produced. When the intensity of the driving force is low, only signals with low amplitudes and low-frequency peak are produced. Therefore, in small tanks, the signals resemble the tank signal of Fig. 20.1, and the bimodal feature is suppressed since the high-frequency portion of the source cannot be used for a low driving force. As the driving force increases to a moderate level, the low-frequency peak also increases in amplitude, and the high-frequency portion of the signal begins to come into use. As the driving force increases further the amplitude of the high-frequency peak becomes larger than that of the low-frequency peak, resulting in type III signals. As the driving force continues to increase to a high level, the amplitude of the high-frequency peak becomes much greater than the amplitude of the low-frequency peak and completely dominates the low-frequency peak causing the bimodal feature to be suppressed. Recent field measurements of free-ranging dolphins [20.55–57] suggest that the majority of echolocation clicks emitted by dolphins are bimodal.

The largest odontocete species is the sperm whale (*Physeter catodon*) and it too emits echolocation signals. Prior to the late 1990s the most prevalent understanding of sperm whale signals is that the clicks were broadband with peak frequencies between 4 and 8 kHz [20.58–61]. There were only two reports on source levels. *Dunn* [20.62] using sonobouys measured 148 sperm whale clicks from a solitary sperm whale and estimated an average peak-to-peak source level of 183 dB re 1 μPa. *Levenson* [20.59] estimated peak-to-peak source level of 180 dB re 1 μPa. The clicks were also thought to be essentially nondirectional and projected in codas [20.63]. Therefore, the notion of sperm whales echolocating was somewhat questionable. Part of the problem was the lack of measurements of sperm whale signaling in conjunction with foraging and the fact that click signals can have more than one function, such as communications and echolocation. In a review paper *Watkins* [20.63] spelled out his rational for not supporting the notion of a sonar function for sperm whale clicks, "Other features of their sounds however, do not so easily fit echolocation:" Watkin's rational included his observations that clicks do not appear to be very directional, the interclick interval does not varied as if a prey or obstacle is being approached, solitary sperm whales are silent for long periods, the level of their clicks appears to be generally greater than that required for echolocating prey or obstacles and the individual clicks are usually too long for good range resolution. It is important to state that most of Watkins measurements were conducted at low-temperate latitudes where females and calves are found.

Sperm whale echolocation began to be more fully understood with data obtained in ground-breaking work by Bertel Møhl and his students from the University of Aarhus, along with other Danish colleagues. They began to perform large-array aperture measurements of large bull sperm whales foraging along the slope of the continental shelf off Andenes, Norway beginning in the summer of 1977 [20.64]. Up to seven multi-platforms spaced on the order of 1 km apart were used in their study with hydrophones placed at depths varying from 5 to 327 m [20.65]. They also came up with a unique but logistically simple scheme of obtaining global positioning system (GPS) information to localize the position of each platform. Each platform continuously logged its position and time stamps on one track of a digital audio tape (DAT) recorder, the other track being used for measuring sperm whale clicks [20.64]. The GPS signals were converted to an analog signal by frequency-shift keying (FSK) modulation. In this way, each platform can be operated essentially autonomously and yet its lo-

Amazon EX1014 Page 822

cation and time stamps could be related to all the other platforms.

An important finding of *Møhl* et al. [20.64] is the monopulsed nature of on-axis clicks emitted by the sperm whales that were similar in shape and spectrum to dolphin echolocation signals but with peak frequencies between 15 and 25 kHz (this peak frequency range is consistent with the auditory sensitivity observed from evoked potential responses by a young sperm whale to clicks presented by *Ridgway* and *Carder* [20.38]). *Møhl* et al. [20.64] also estimated very high source levels as high as 223 dB re 1 μPa per RMS (root mean square). The per RMS measure is the level of a continuous sine wave having the same peak-to-peak amplitude as the click. This measure is clearly an overestimate of the RMS value of a sperm whale click. Nevertheless, the 223 dB reported by *Møhl* et al. [20.64] can easily be given a peak-to-peak value of 232 dB (adding 9 dB to the per-RMS value). In a follow-on study, *Møhl* et al. [20.65] measured clicks with RMS source levels as high as 236 dB re 1 μPa using the expression

$$p_{\mathrm{RMS}} = \sqrt{\frac{1}{T}\int_0^T p^2(t)\,\mathrm{d}t}\,. \tag{20.5}$$

For this measurement, they used a time interval corresponding to the 3 dB down points of the signal waveform. For this waveform, an RMS source level of 236 dB corresponds to a peak-to-peak source level of 243 dB re 1 μPa. Finally, they found that the clicks were very directional, as directional as dolphin clicks.

20.2.2 Echolocation Signals of Smaller Odontocetes that Do not Whistle

The second class of echolocation signals are produced by dolphins and porpoises that do not emit whistle signals. Not many odontocetes fall into this category and these non-whistling animals tend to be smaller than their whistling cousins. Included in this group of non-whistling odontocetes are the harbor porpoise (*phocoena phocoena*), finless porpoise (*neophocaena phocaenoides*), Dall's porpoise (*phocoenoides dalli*), Commerson's dolphin, *cephalorhynchus commersonii*), Hector's dolphin (*cephalorhynchus hectori*) and pygmy sperm whale (*kogia sp*).

Examples of harbor porpoise and Atlantic bottlenose dolphin echolocation signals presented in a manner for easy comparison are shown in Fig. 20.9. There are four fundamental differences in the two types of echolocation signals. The non-whistling animal emit a signal with longer duration, narrower band, lower amplitude and single mode. The length of the *Phocoena phocoena* signal vary from about 125–150 μs compared to 50–70 μs for the *tursiops truncatus* signal. The bandwidth of the *Phocoena* signal is almost 0.2–0.3 that of the *tursiops* signal. Since the non-whistling dolphins are usually much smaller in length and weight than the whistling dolphins, the smaller animals might be amplitude limited in terms of their echolocation signals and compensate by emitting longer signals to increase the energy output. This issue can be exemplified by characterizing an echolocation click as [20.39]

$$p(t) = As(t)\,, \tag{20.6}$$

where $A = |p|_{\max}$ is the absolute value of the peak amplitude of the signal and $s(t)$ is the normalized wave-

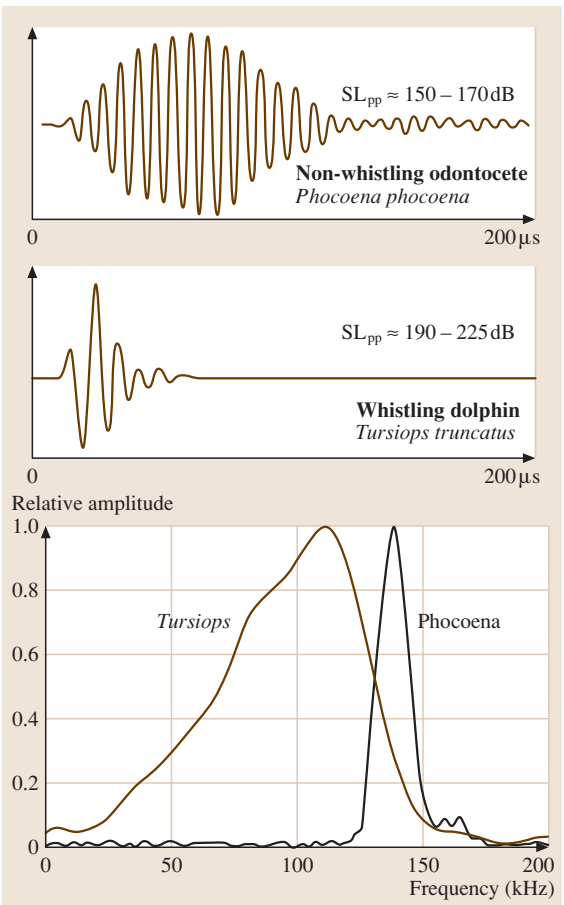


Fig. 20.9 Examples of typical echolocation signals of *phocoena phocoena* and *tursiops truncatus* (after [20.8])

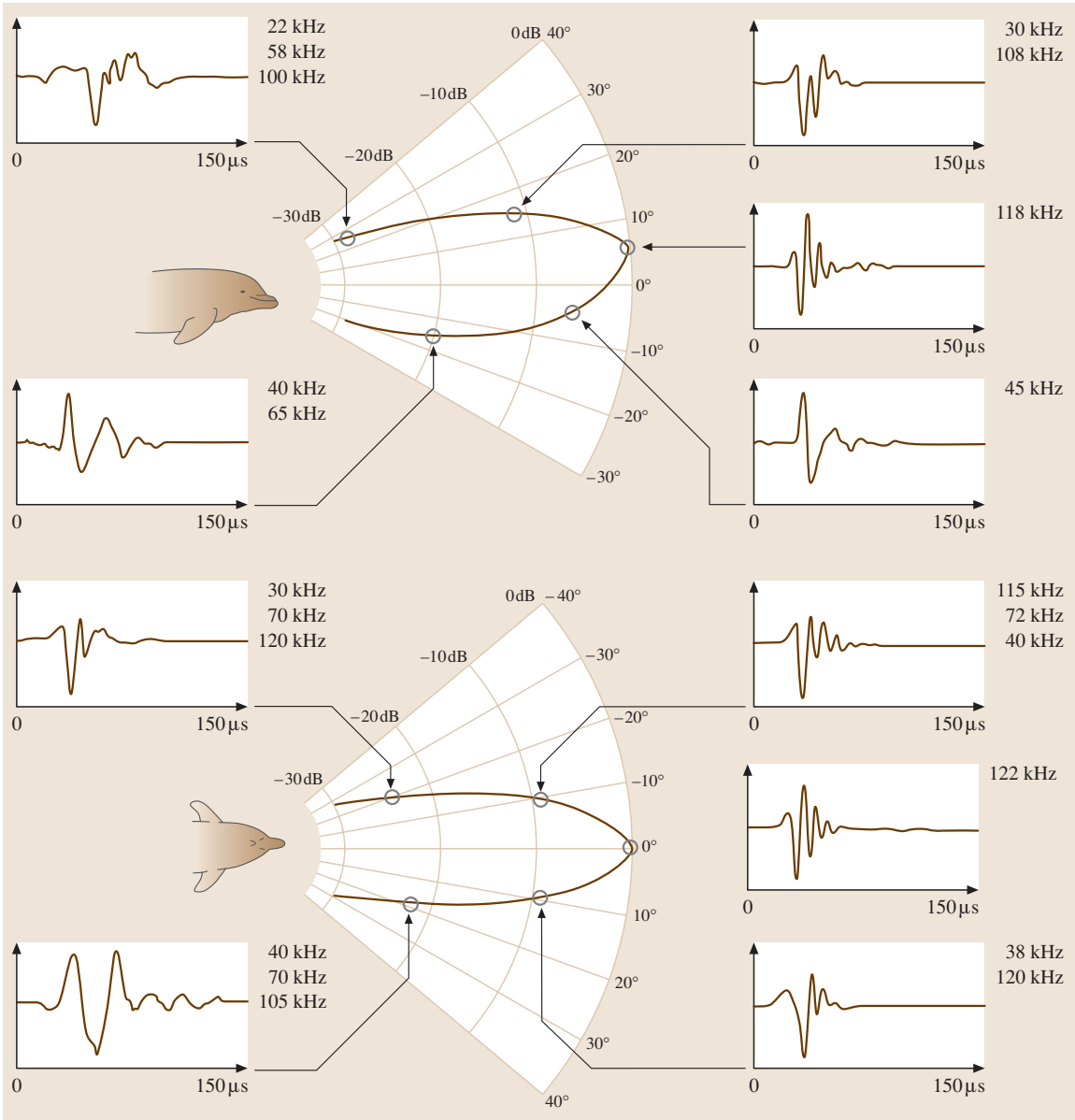


Fig. 20.10 The transmission beam pattern of a *tursiops truncatus* planes with the waveform of a click measured by 5–7 hydrophones (after *Au* [20.39]) in the vertical and horizontal planes

form having a maximum amplitude of unity. The source energy-flux density of the signal in dB can be expressed as

$$\mathrm{SE} = 10\log\left(\int_0^T p(t)^2\,\mathrm{d}t\right)$$

$$= 10\log(A) + 10\log\left(\int_0^T s(t)^2\,\mathrm{d}t\right), \qquad (20.7)$$

where T is the duration of the signal. Letting $2A \approx$ be the peak-to-peak sound pressure level, (20.4) can be

rewritten as

$$\mathrm{SE} = \mathrm{SL} - 6 + 10\log\left(\int_0^T s(t)^2\,\mathrm{d}t\right)\,, \qquad (20.8)$$

where SL is the the peak-to-peak source level and is approximately equal to $10\log(2A)$. From Fig. 20.9, the *Tursiops* can emit signals with peak-to-peak source levels in open water that are greater than 50 dB for *phocoena*. The portion of the source energy that can be attributed to the length of the signal is only about 2–4 dB greater for *phocoena* than *Tursiops* [20.39]. Therefore, one can conclude that the target detection range of *phocoena* is considerably shorter than for *Tursiops*. This has been demonstrated by *Kastelein* et al. [20.67] who measured a target detection threshold range of 26 m for a 7.62 cm water-filled sphere in a quiet environment. The target detection range in a noisy environment and a similar 7.62 cm water-filled sphere was 113 m [20.68].

The echolocation signals of other non-whistling odontocetes are very similar to the *Phocoena phocoena* signal shown in Fig. 20.9. Examples of the echolocation signals of other non-whistling odontocetes can be found in *Au* [20.7, 39, 69]. Unfortunately, reliable source-level data have been collected only for *Phocoena phocoena* [20.70] and *Kogia* [20.69].

20.2.3 Transmission Beam Pattern

The outgoing echolocation signals of dolphins and other odontocetes are projected in a directional beam that have been measured in the vertical and horizontal planes for the bottlenose dolphin [20.39], beluga whale [20.66] and the false killer whale [20.52]. The composite beam pattern from the three measurements on *Tursiops* along with the averaged waveform from a single trial measured by 5 or the 7 hydrophones are shown in Fig. 20.10. The 3 dB beam width for the bottlenose dolphin was approximately 10.2° in the vertical plane and 9.7° in the horizontal plane. The waveforms detected by the various hydrophones in Fig. 20.10 indicate that signals measured away from the beam axis will be distorted with respect to the signals measured on the beam axis. The further away from the beam axis the more distorted the signals will be. This distortion come from the broadband nature of the click signals emitted by whistling dolphins. This characteristics also make it difficult to get good measurements of echolocation signals in the field even with an array of hydrophones since it is extremely difficult to obtain on-axis echolocation signals and also to know the orientation of the animal with respect to the measuring hydrophones. The frequencies shown with each waveform are the fre-

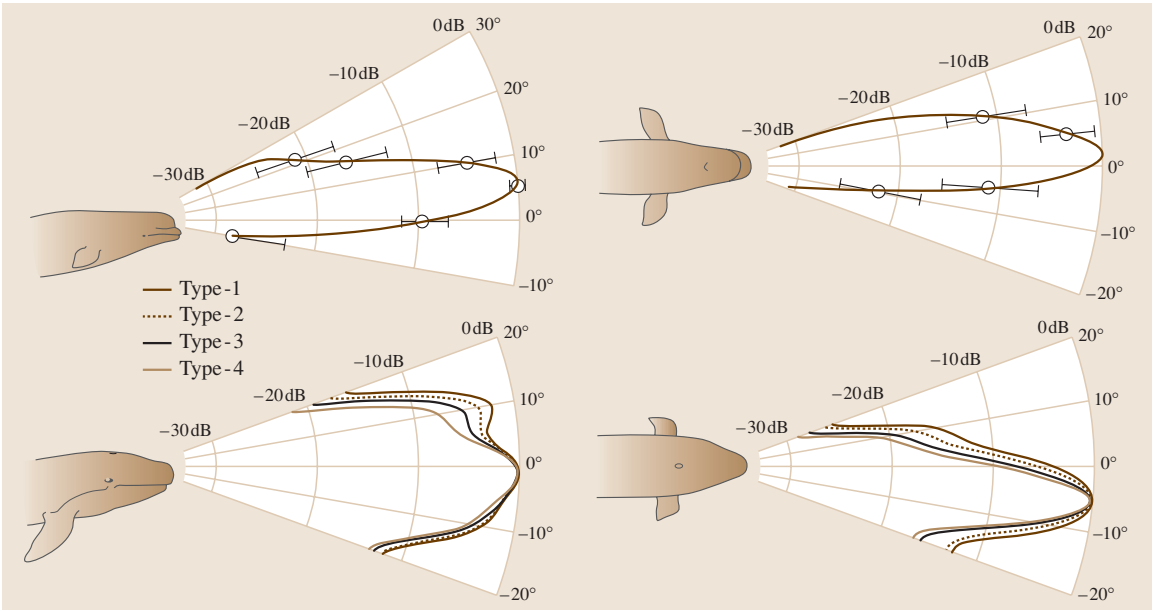


Fig. 20.11 The transmission beam pattern of a *delphinapterus leucas* (after *Au* et al. [20.66]) and a *pseudorca crassidens* (after *Au* et al. [20.52]) in the vertical and horizontal planes

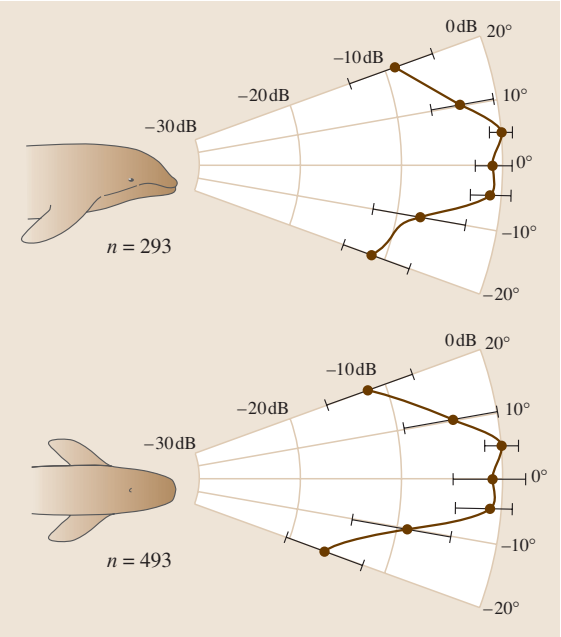


Fig. 20.12 Beam patterns in the vertical and horizontal planes for *phocoena phocoena* (*Au* et al. [20.70])

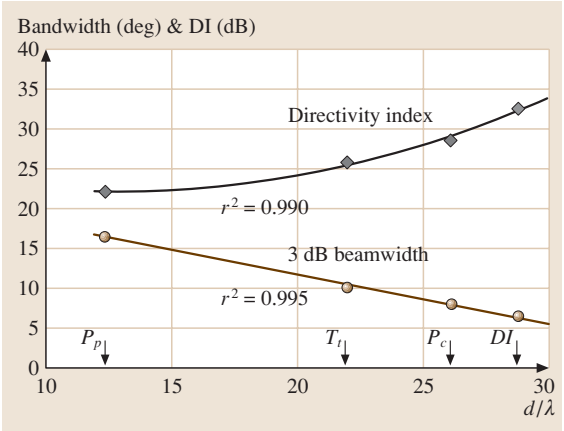


Fig. 20.13 Transmission directivity index and 3 dB beam width for four odontocetes. The directivity index and beam width for *Tursiops truncatus* and *delphinapterus leucas* came from *Au* [20.39] and for *pseudorca crassidens* after *Au* et al. [20.52]. The wavelength λ corresponds to the average peak frequency of the animals' echolocation signals. The directivity index is fitted with a second-order polynomial curve the the beam width is fitted with a linear curve (after *Au* et al. [20.70])

quencies of local maxima in the spectrum. At some angles the averaged signal had multiple peaks in the spectrum and these peaks are listed in order of prominence.

The beam patterns for the beluga and false killer whale are shown in Fig. 20.11. The beam width for the beluga whale was approximately 6.5° in both planes. Four beam patterns corresponding to the four signal types described in Fig. 20.7 are shown for the false killer whale. For the highest frequency (type IV signal) the 3 dB beam width in the vertical plane was approximately 10°, and 7° in the horizontal plane. The beam axis in the vertical plane for the bottlenose dolphin and the beluga whale was +5° above the horizontal plane. For the false killer whale, 49% of the beam axis was at 0° and 32% at −5°. The four beam patterns for the false killer whale indicate that, like a linear transducer, the lower the frequency the wider the beam pattern. The type I signal has a peak frequency of about 30 kHz and has the widest beam. However, even though the peak frequency of the type IV signal is about 3.5 times higher than the type I signal, the beam does not seem to be substantially larger. This property can be used when discussing lower-frequency whistle signals to suggest that whistle signals are also emitted in a beam.

The only transmission beam pattern for a non-whistling odontocete, a *phocoena phocoena*, was measured by *Au* et al. [20.70]. Their results are shown in Fig. 20.12 in both the horizontal and vertical planes. One of the obvious difference between the beam patterns in Fig. 20.12 and those in Figs. 20.10 and 20.11 is the width of the beam. Although the harbor porpoise emitted the highest-frequency signals, its small head size caused the beam to be wider than for the other animals. The beam patterns of Figs. 20.10–20.12 were inserted into (20.1) and numerically evaluated to estimate the corresponding directivity index for *tursiops* and *delphinapterus* [20.39] and for *pseudorca* by *Au* et al. [20.52] and for *phocoena* [20.70], and the results are shown in Fig. 20.13 plotted as a function of the ratio of the head diameter (at the blow hole) of the subject and the average peak frequency of the animals echolocation signal. Also shown in Fig. 20.13 are the 3 dB beam width, where the 3 dB beam width in the vertical and horizontal planes were averaged.

The second-order polynomial fit of the directivity index is given by the equation

$$\mathrm{DI} = 28.89 - 1.04\left(\frac{d}{\lambda}\right) + 0.04\left(\frac{d}{\lambda}\right)^2 . \qquad (20.9)$$

The linear fit of the 3 dB beam width data is given by the equation

$$\mathrm{BW} = 23.90 - 0.60\left(\frac{d}{\lambda}\right) . \qquad (20.10)$$

The results of Fig. 20.13 provides a way of estimating the directivity index and 3 dB beam width of other odontocetes by knowing their head size and the peak frequency of their echolocation signals.

Another interesting characteristics of the beam pattern measurement on the *phocoena phocoena* is that the off-axis signals are not distorted in comparison to the on-axis signal. This is consistent with narrow-band signals whether for a linear technological transducer or a live animal. Therefore, measurements of echolocation signals of non-whistling odontocetes in the field can be performed much easier than for whistling odontocetes.

20.3 Odontocete Acoustic Communication

Members of the *Odontocete* suborder are an intriguing example of the adaptability of social mammalian life to the aquatic habitat. Most odontocetes, particularly marine dolphins, live in groups ranging from several to hundreds of animals in size, forage cooperatively, develop hierarchies, engage in alloparental care and form strong pair bonds and coalitions between both kin and non-kin alike (see [20.71–73] for a review of the literature on dolphin societies). These social traits are analogous to patterns found in many avian and terrestrial mammalian species. It is not surprising, therefore, that odontocetes mediate much social information via communication. Some river dolphins are usually found as solitary individuals or mother–calf pairs although they may occasionally congregate into larger groups.

Odontocetes communicate through a combination of sensory modalities that include the visual, tactile, acoustic and chemosensory channels [20.74]. Visual signals in the form of postural displays are thought to convey levels of aggression, changes in direction of movement, and affiliative states [20.72,75]. Tactile interactions vary in purpose from sexual and affiliative signals to expressions of dominance and aggression [20.76,77]. Although not well documented yet, it is thought that a derivative of chemosensory perception, termed *quasiolfaction* allows dolphins to relay chemical messages about stress and reproductive status [20.78, 79]. However, it is the acoustic communication channel in particular that is believed to be the main communication tool that enables odontocetes to function cohesively as groups in the vast, visually limited environment of the open sea [20.80]. Examining how odontocetes have adapted their use of sound is therefore an important step to understanding how social marine mammal life evolved in the sea and how it has found a way to thrive in a habitat so drastically different from our own.

When compared to all that has been learned about dolphin echolocation over the past several decades, surprisingly little is still known about how dolphins and other odontocetes use acoustic signals for communication. A major reason for this is that it is very difficult to observe the acoustic and behavioral interactions between the producer and the receiver(s) of social signals in the underwater environment. Sound propagation in the sea makes even a simple task like identifying the location of a sound source very challenging for air-adapted listeners. Therefore, matching signals with specific individuals and their behavior in the field is problematic without the use of sophisticated recording arrays [20.57, 76, 81, 82], or directional transducers. In addition, many questions remain unanswered about the nature of social signals themselves. Much still remains unknown about the functional design of dolphin social sounds. This is largely due to the fact that most species specialize in producing and hearing sounds with frequencies well beyond the limits of the human hearing range. Limitations in the technology available to study social signaling in the field have until recently restricted most analyses to the human auditory bandwidth (< 20 kHz). Yet, despite these significant challenges, a great deal of progress has been made in our understanding of odontocete acoustic communication. Rapidly advancing technologies are contributing greatly to our ability to study social signaling both in the field and in the laboratory. The emerging picture reveals that odontocetes have adapted their acoustic signaling to fit the aquatic world in a remarkably elegant way.

20.3.1 Social Acoustic Signals

Odontocetes have evolved in both marine and freshwater habitats for over 50 million years. Over such a long period of adaptive radiation one might expect that a variety of signaling strategies would have evolved among the approximately 65 species of small toothed whales and dolphins. Yet, remarkably, the vast majority of these species have either conserved or converged on the pro-

duction of two types of sounds: whistles and short pulse trains.

Whistles

Whistles are frequency- and amplitude-modulated tones that typically last anywhere from 50 ms to 3 s or more, as illustrated in Fig. 20.14. They are arguably the most variable signals produced by dolphins. Both within and across species they range widely in duration, bandwidth and degree of frequency modulation. How whistles are used in communication is an ongoing topic of debate among researchers, but most agree that they play an important role in maintaining contact between dispersed individuals and are important in social interactions [20.85].

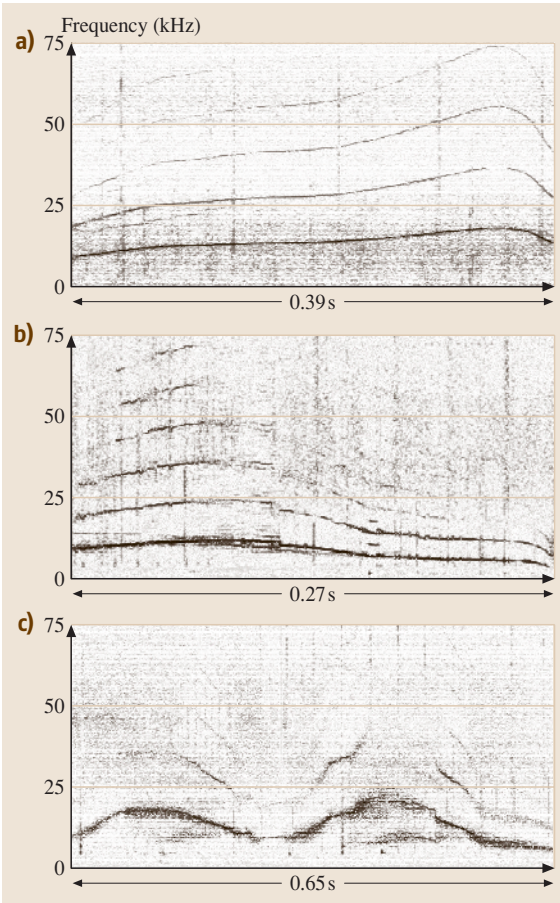


Fig. 20.14a–c Variations in whistle forms produced by Hawaiian spinner dolphins **(a)** and Atlantic spotted dolphins **(b)**, **(c)** exhibiting multiple harmonics (after *Lammers* et al. [20.83, 84])

Dolphin whistles exhibit both species and geographic specificity [20.86–90]. Differences are greatest between distantly related taxa and between species of different size. As a general rule, larger species tend to produce whistles with lower fundamental frequencies than smaller ones [20.88]. Geographic variations within a species are usually smaller than interspecific differences [20.87], but some species do develop regional dialects between populations [20.89, 90]. These dialects tend to become more pronounced with increasing geographic separation. In addition, pelagic species tend to have whistles in a higher-frequency range and with more modulation than coastal and riverine species [20.86, 88]. Such differences have been proposed as species-specific reproductive-isolating characteristics [20.86], as ecological adaptations to different environmental conditions and/or resulting from restrictions imposed by physiology [20.88]. On the other hand, it must also be noted that numerous whistle forms are also shared by both sympatric and allopatric species [20.91]. The communicative function of shared whistle forms is unknown, but it is intriguing as it suggests a further tendency towards a convergence in signal design.

Dolphins produce whistles with fundamental frequencies usually in the human audible range (below 20 kHz). However, whistles typically also have harmonics (Fig. 20.14), which occur at integer multiples of the fundamental and extend well beyond the range of human hearing [20.84]. Harmonics are integral components of tonal signals produced by departures of the waveform from a sinusoidal pattern. Dolphin whistle harmonics have a *mixed directionality* property, which refers to the fact they become more directional with increasing frequency [20.83, 92]. It has been proposed that this signal feature functions as a cue, allowing listening animals to infer the orientation and direction of movement of a signaling dolphin [20.83, 92]. Harmonics may therefore be important in mediating group cohesion and coordination.

Burst Pulses

In addition to whistles, most odontocetes also produce pulsed social sounds known as *burst pulse* signals. Burst pulse signals are broadband click trains similar to those used in echolocation but with inter-click intervals of only 2–10 ms [20.93]. Because these intervals are considerably shorter than the processing period generally associated with echolocation and because they are often recorded during periods of high social activity, burst pulse click trains are thought instead to play an important role in communication.

Amazon EX1014 Page 828

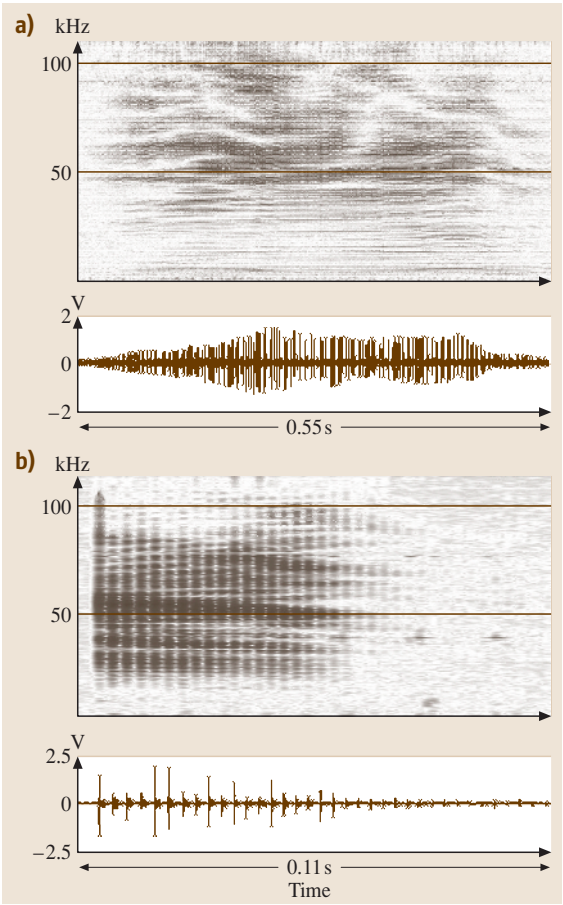


Fig. 20.15a,b Examples of **(a)** high-quantity and **(b)** low-quantity burst pulses produced by Atlantic spotted dolphins (*stenella frontalis*). Click train *a* has 255 clicks with mean ICI (interclick interval) of 1.7 ms. Click train *b* has 35 clicks with a mean ICI of 2.9 ms (after *Lammers* et al. [20.83, 84])

Burst pulses vary greatly in the inter-pulse interval and in the number of clicks that occur in a train, which can number anywhere from three to hundreds of pulses, as depicted in Fig. 20.15. This variation gives them distinctive aural qualities. Consequently, burst pulse sounds have been given many subjective labels, including *yelps* [20.94], *cracks* [20.95], *screams* [20.96] and *squawks* [20.97]. Their production has been reported in a variety of odontocete species including: the Atlantic bottlenose dolphin (*tursiops truncatus*) [20.95], the Hawaiian spinner dolphin (*stenella longirostris*) [20.72], the Atlantic spotted dolphin (*stenella frontalis*) [20.98], the narwhal (*monodon monoceros*) [20.99], Hector's dolphin (*cephalorhynchus hectori*) [20.100], the pilot whale (*globicephala melaena*) [20.101], and the harbor porpoise (*phocoena phocoena*) [20.101]. To date, much remains unknown about how burst pulses function as communication signals. It is generally believed that they play an important role in agonistic encounters because they are commonly observed during confrontational head-to-head behaviors between individuals [20.95, 102–104]. However, some authors have suggested they may represent emotive signals in a broader sense [20.97, 105, 106], possibly functioning as graded signals [20.107].

Given that dolphins have temporal discrimination abilities well within the range required to resolve individual clicks in a burst pulse [20.39, 108], it is possible that the quantity of clicks and their temporal spacing could form an important basis for communication. However, as with whistles, no data presently exist on the classification and discrimination tendencies of dolphins with respect to different burst pulses.

20.3.2 Signal Design Characteristics

Although much remains unknown about how odontocetes use acoustic signals for communication, the communicative potential of their signals can, in part, be inferred by considering the design characteristics that have been uncovered thus far. Signal features such as their detectable range, the production duty cycle, the identification level, the modulation potential and the form–content linkage provide useful clues about how odontocetes might use whistles and burst pulses.

The Active Space of Social Signals

The effective range of signals used for communication is generally termed the *active space*. *Janik* [20.109] investigated the active space of whistles produced by bottlenose dolphins in the Moray Firth, Scotland, using a dispersed hydrophone array to infer geometrically the location of signaling animals and establish the source level of their whistles. The mean source level was calculated to be 158 dB ± 0.6 re 1 µPa, with a maximum recorded level of 169 dB. By factoring in transmission loss, ambient noise levels, the critical ratios and auditory sensitivity of the species involved the active space of an unmodulated whistle between 3.5 and 10 kHz in frequency was estimated to be between 20 and 25 km in a habitat 10 meters deep at sea state 0. At sea state 4 the estimated active space ranged from 14 to 22 km, while for whistles at 12 kHz it dropped to between 1.5 and 4 km. These estimations

Amazon EX1014 Page 829

were made for dolphins occurring in a relatively shallow, mostly mud-bottom, quiet environment. Presently, no data exist on the active space characteristics of delphinid whistles in pelagic waters and comparatively louder tropical near-shore environments. Similarly, no estimates have yet been made for the active space of burst pulses.

The Duty Cycle of Social Signals

Duty cycle refers to the proportion of time that any given signal or class of signals is *on* versus *off*. Signals can vary in the fine structure of their temporal patterning (e.g. duration), in their temporal spacing within a bout, and in their occurrence within a larger cyclical time frame, such as a 24 hour day. Each of these aspects of the temporal delivery of signals carries its own implication for communication. Fine-scale characteristics can define the nature of the signal itself and provide information about its relationship to other signals as well as constraints associated with its mechanism of production. The occurrence and timing of a signal within a bout can help convey information about the urgency of a situation, the level of arousal, the fitness of an individual and can assist the receiver in the task of localization. Finally, the periodicity of signals over hours or days can be an indicator of variables such as activity levels and reproductive state.

Murray et al. [20.110] investigated the fine-scale duty-cycle characteristics of delphinid social signals. Using a captive false killer whale (*pseudorca crassidens*) as their subject, they examined the temporal relationship between the units of a click train (individual clicks) and tonal signals. Their analysis revealed that false killer whales modulate the temporal occurrence of clicks to the point of grading them into a continuous wave (CW) signal such as a whistle. This finding was interpreted as evidence that *pseudorca* may employ graded signaling for communication (a topic discussed in more detail below), as well as to suggest that clicks and whistles are produced by the same anatomical mechanism.

The timing of signals within a bout has not been investigated with much success among delphinids. The primary obstacle towards this line of work has been the difficulty of identifying the signaler(s) involved for even short periods of time under field conditions. Early work by *Caldwell* and *Caldwell* [20.111] on a group or four naïve captive common dolphins (*delphinus delphis*) suggests that whistle exchanges do have temporal structure. In whistling bouts involving more than one signaler, the onset of a whistle specific to an individual was followed within two seconds by that of another. Furthermore, initiation of a whistle by two animals within 0.3 s of one another always resulted in the inhibition of one of them, with some individuals deferring more than others. Initiations separated by 0.4–0.5 s caused inhibition less frequently while those longer than 0.6 s resulted in almost no inhibition. Repeated whistles produced without an intervening response by another animal were usually delayed by less than one second. Thus, the duty cycle of whistling bouts and chorusing behavior among common dolphins does appear to follow certain temporal rules, but their significance is not clear.

Periodicity in social signaling has been investigated in captive bottlenose dolphins and common dolphins [20.112–114], as well as free-ranging spinner dolphins (*stenella longirostris*) [20.72] and common dolphins [20.115]. In the captive studies, signaling activity was linked to feeding schedules, nearby human activities and responses to different forms of introduced stress [20.113]. For spinner and common dolphins in the wild, the occurrence of social acoustic signals was highest at night, when both species were foraging, and lowest in the middle of the day.

The Identification Level of Social Signals

The role of delphinid whistles as individual-specific signals has been the focus of more scientific attention than any other aspect of their social acoustic signaling. *Caldwell* and *Caldwell* [20.116] were the first to propose that individual dolphins each possess their own distinct *signature* whistle. The idea was borne out of the observation that recently captured dolphins each produce a unique whistle contour that makes up over 90% of the individual's whistle output. Since being proposed, the so-called *signature whistle hypothesis* has emerged as the most widely accepted explanation for whistling behavior among dolphins. The idea has received support from numerous studies involving captive and restrained animals [20.85, 117–121] as well as from field studies of free-ranging animals [20.97, 122, 123]. Some, however, have argued that a simple signature function for dolphin whistling cannot account for the diversity of signals observed in socially interactive dolphin groups [20.124, 125]. While not denying the presence of signature whistles per se, these authors have argued that the large percentage of stereotyped signature signals observed in other studies may be an artifact of the unusual circumstances under which they were obtained (isolation, temporary restraint, separation, captivity). The debate over the prevalence of signature whistles among captive and free-ranging dolphins remains a contested topic in the

literature and at scientific meetings. On the other hand, no evidence or formal discussions presently exist suggesting burst pulse signals carry any individual-specific information.

The Modulation Potential of Social Signals

The modulation potential describes the amount of variation present in any given signal as measured by its position along a scale from stereotyped to graded signaling [20.126]. Stereotyped signals are repeats of structurally identical forms and vary discretely between one another. They are used most often for communication when the prospect of signal distortion is high, such as in noisy or reverberant environments or for communicating over long ranges. Graded signals have more variants than stereotyped ones and are encoded by changing one or more signal dimensions. These can include intensity, repetition rate and/or frequency and amplitude modulation. Graded signals are usually employed when a continuous or relative condition must be communicated in a favorable propagation environment.

Many studies to date have made an a priori assumption that dolphin signaling is categorical in nature, with signals belonging to mutually exclusive classes on the basis of their shared similarities. Some evidence in support of this assumption comes from the signature-whistle hypothesis work, where restrained and isolated individuals are often observed producing highly stereotyped bouts of signaling. However, few data are presently available to suggest how dolphins perceive and distinguish social signals [20.127]. Therefore, it is unclear whether an assumption based on the occurrence of signature whistles is broadly applicable towards other forms of social acoustic signaling (i. e. non-signature whistles and burst pulses).

A few studies have explored the occurrence of graded signaling in the communication of delphinids and the larger whales. *Taruski* [20.128] created a graded model for the whistles produced by North Atlantic pilot whales (*globicephala melaena*). He concluded that these signals could be arranged as a "continuum or matrix (of signals) from simple to complex through a series of intermediates" (*Taruski* [20.128] p. 349). *Murray* et al. [20.110], examining the signals of a captive false killer whale (*Pseudorca crassidens*), came to a similar conclusion and proposed that *Pseudorca* signals were also best represented along a continuum of signal characteristics, rather than categorically. Finally, *Clark* [20.129], examining the "call" signals of southern right whales (a mysticete), made the observation that " ... the total repertoire of calls is best described as a sequence of intergraded types along a continuum" (*Taruski* [20.128] p. 1066). Therefore, graded signaling may well play a role in odontocete communication. Controlled perceptual experiments and a better understanding of the acoustic environment dolphins inhabit are needed to test how odontocetes discriminate and classify social sounds.

The Form–Content Linkage of Social Signals

The degree to which the form of a signal is linked to its content depends on how information is coded and on proximate factors associated with the signal's production. The relationship between a signal's structural form and its message can range from being rather arbitrary to tightly linked to a specific condition [20.126]. Among odontocetes, a clear relationship between signal form and content has only been demonstrated for signature whistles (discussed above). Dolphins have been experimentally shown to be capable of labeling objects acoustically (also known as vocal learning), as well as mimicking model sounds following only a single exposure [20.130]. Additional evidence of vocal learning exists from culturally transmitted signals in the wild [20.131] and in captivity [20.132], as well as from the spontaneous production of acoustic labels in captivity [20.127]. Context-specific variations in signature whistle form have also been demonstrated [20.121]. However, to date, no semantic rules have been identified for naturally occurring social signals among odontocetes.

Geographic Difference and Dialect

There is a distinct difference between geographical difference and dialect. Geographical differences are associated with widely separated populations that do not normally mix. Dialect is best reserved for sound emission differences on a local scale among neighboring populations which can potentially intermix [20.133]. Geographic variations are generally considered to result from acoustic adaptations to different environments, or a functionless byproduct of isolation and genetic divergence caused by isolation [20.134]. The functional significance of dialects is controversial, with some maintaining that dialects are epiphenomena of song learning and social adaptation, whereas others believe that they play a role in assorted mating and are of evolutionary significance [20.134]. Dialects are known to occur in many species of birds [20.135] but appear to be very rare in mammals.

Amazon EX1014 Page 831

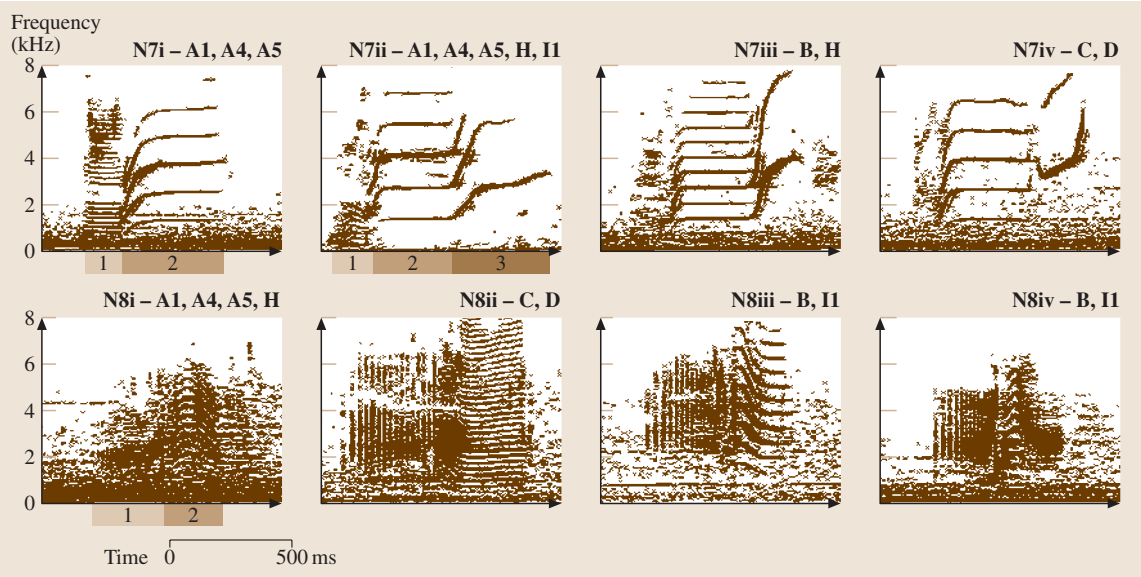


Fig. 20.16 Spectrograms of call types N7 and N8 for clan A. Above each spectrogram is the subtype identification and the pods that produce the variant, and below certain spectrograms are division marks separating calls into their component parts (after *Ford* [20.134]).

Killer Whale

Killer whales (*orcinus orca*) produce a specific type of burst pulse termed the *discrete call*. Discrete calls are thought to serve as contact calls between individuals, much like whistles in other odontocete species [20.136]. Discrete calls are population specific and even pod specific. The dialects of resident killer whales (*Orcinus orca*) in the coastal waters of British Columbia and Washington have been studied over a prolonged period by *Ford* and his colleagues [20.134,137,138]. They used photographic identification techniques, keying on unique natural markings on the whales to identify 16 pods or stable kin groups of 232 resident killer whales. Differences in acoustic behavior formed a system of related pod-specific dialects within the call tradition of each clan. *Ford* [20.134] has proposed that each clan is comprised of related pods that have descended from a common ancestral group and pod-specific repertoires probably serve to enhance the efficiency of acoustic communications within the group and act as behavioral indicators of pod affiliation.

Killer whale calls are typically made up of rapidly emitted pulses that to the human ear have a tonal quality [20.134]. Many calls have several abrupt shifts in pulse repetition rate allowing them to be divided into different segments or parts. Although all pods belonging to a clan share a number of calls, these calls were often rendered in consistently different form by different pods. Also, certain pods produced calls that were not used by the rest of the clan. Such variations produced a set of related group-specific dialects within the call tradition of each clan.

Spectrograms of call types N7 and N8 are shown in Fig. 20.16. In this example, we can see that different pods produce similar but different versions of call type N7 and N8. Pods A1, A4 and A5 produced two versions of call type N7 and pod B and I1 produced two versions of call type N8. All of the spectrograms have two parts except for call type N7ii, which had three parts.

Ford [20.134] contended that pod-specific repertories can be retained for periods in excess of 25 years. Discrete calls generally serve as signals for maintaining contact within the pod and that the use of repertoire of pod-specific calls enhances this function by conveying group identity and affiliation [20.134]. Killer whales, like most other small dolphins are able to learn and mimic a wide variety of sounds. Even from a young age, killer whale infants can selectively learn specific calls, especially the calls of their mothers. *Bowles* et al. [20.139] studied the development of calls of a captive-born killer whale calf and found that it learned and reproduced only the calls of its mother and ignored the calls of other killer whales in the same pool. *Ford* [20.134] also observed killer whales imi-

Amazon EX1014 Page 832

tating the call types of different pods, and even those from other clans. These instances were rare, but it does show a capacity for learning and mimicry of acoustic signals.

Sperm Whales

Sperm whales live in a matrilineal family unit where there exist cooperative behaviors including communal care of the young in ways similar to killer whales. The family units are very stable and females may live as long as 60–70 y [20.140]. One type of signals that sperm whales emit are denoted as codas, which are sequences of click signals that may be the primary means of acoustic communications for these animals [20.58]. *Weilgart* and *Whitehead* [20.141] recorded, for over a year, codas from a number of sperm whales around the South Pacific and in the Caribbean Sea. Photographic identification also allowed them to assign recording sessions to particular groups. They found that the coda repertoire recorded from the same group on the same or different days were much more similar than those recorded from different groups in the same place. Groups recorded in the same place had more similar coda repertoire than those in the same broad area but different places. Groups from the same area were in turn marginally similar to those in the same ocean but different than those in different oceans. Coda class repertoires of groups in different oceans and in different areas within the same ocean were statistically significantly different. They concluded that strong group-specific dialects were apparently overlaid on weaker geographic variation. Sperm whales, killer whales and possibly bottlenose dolphins are the only cetaceans known to have dialects.

20.4 Acoustic Signals of Mysticetes

There are eleven species of mysticetes or baleen whales and sounds have been recorded from all but the pygmy right whale [20.142]. The vocalization of baleen whales can be divided into two general categories: (1) songs and (2) calls [20.142]. The calls can be further subdivided into three categories: (1) simple calls, (2) complex calls and (3) clicks, pulses, knocks and grunts [20.142]. Simple calls are often low-frequency, frequency-modulated signals with narrow instantaneous bandwidth that sound like moans if a recording is speeded up or slowed down, depending on the specific animal. Amplitude modulation and the presence of harmonics are usually part of a simple call, with most of the energy below 1 kHz. Complex calls are pulse-like broadband signals with a variable mixture of amplitude and/or frequency modulation. They sound like screams, roars, and growls, with most of the energy between 500–5000 Hz. Clicks, pulses, knocks and grunts are short-duration (< 0.1 s) signals with little or no frequency modulation. Clicks and pulses are very short (< 2 ms) signals with frequencies between 3–31 kHz, while grunts and knocks are longer (50–100 ms) signals in the 100–1000 Hz range [20.142].

20.4.1 Songs of Mysticete Whales

Songs are defined as "sequences of notes occurring in a regular sequence and patterned in time" [20.142]. Songs are easily discriminated from calls in most instances. Four mysticetes species have been reported to produce songs; the blue whale (*balaenoptera musculus*) [20.143], the fin whale (*balaenoptera physalus*) [20.144, 145], the bowhead whale (*balaena mysticetus*) [20.15], and the humpback whale [20.146]. Songs of humpback whales have without a doubt received the most attention from researchers. Part of the reason for this is the relative ease for investigators to travel and do research in the summer grounds of humpback whales, especially in Hawaii in the Pacific and Puerto Rico in the Atlantic.

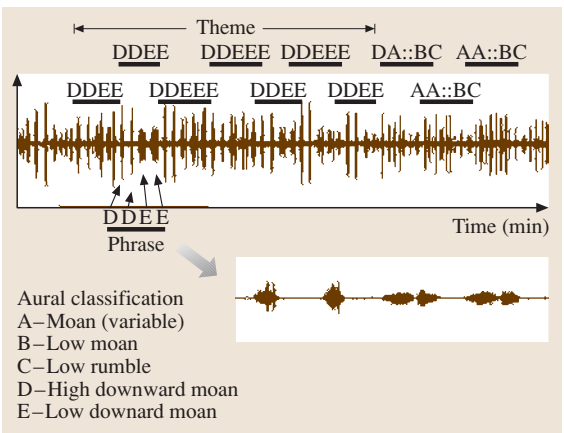


Fig. 20.17 Examples of two themes from a humpback whale song (after *Frankel* [20.36])

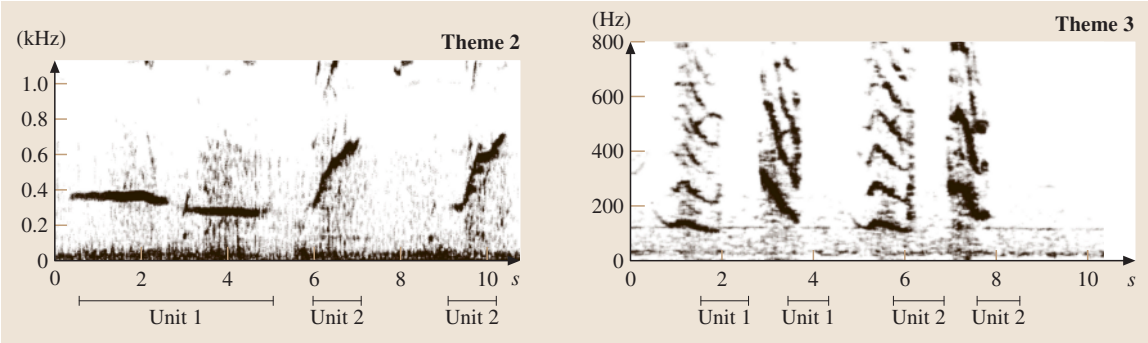


Fig. 20.18 An example of a portion of a humpback whale song. The individual pulse sounds are defined as units and units are aurally categorized based on how they sound. In this example, there are five units. Phrases are formed by a combination of units and themes are formed by a combination of phrases. The whole song consist of a combination of themes that are continuously repeated for the length of the song

Humpback Whale Songs

The list of studies involving humpback whale songs is long and extends from 1971 to the present time, *Helweg* et al. [20.147]. Although some songs may sound almost continuous to human listeners the basic units in a song are pulsed sounds. Songs are sung only by males and consist of distinct units that are produced in some sequence to form a phrase and a repeated set of phrases form a theme and repeated themes form a song. A song can last from minutes to hours depending on the disposition of the singer. An example of a portion of song is shown in Fig. 20.17. Example of two themes of a song in spectrogram format are shown in Fig. 20.18. The variation in frequency as a function of time is clearly shown in the figure.

Some general properties of songs and the whales that are singing are:

1. Songs from the North Pacific, South Pacific and Atlantic populations are different.
2. Singing peaks during the winter months when humpback whales migrate to warmer waters at lower latitudes.
3. Whales within a population sing the same basic song in any one year, although the song may undergo slight changes during a breeding season.
4. Changes in songs are not due to forgetting during the summer months, which are non-singing months, since songs recorded early in the winter breeding season are nearly the same as songs recorded late in the previous breeding season.
5. Songs from consecutive years are very similar but songs across nonconsecutive years will have fewer similarities.
6. Singers are most probably only males, since no females have been observed singing.
7. Some singing also occurs during the summer and fall.
8. Singing whales are often alone, although they have been occasionally observed singing in the presence of other humpback whales.
9. Singers tend to remain stationary. However, they have also been observed singing while swimming.

Other Mysticetes

Bowhead whales emit a variety of different types of simple and complex sounds that sound like moans to the human ear. They also emit sequential sounds that contain repeatable phrases that can be classified as songs during their spring migration [20.15] but not during the summer or autumn [20.148]. The bowhead song had just one theme with basically only two sounds repeated over and over. *Ljungblad* et al. [20.15] reported that songs were very tonal with clear pitch even though they were produced by pulsive moans, whereas *Cummings* and *Holliday* [20.149] described songs as sounding like raucous elephant roars and trumpeting in discrete repetitions or phases that were put together to form longer sequences. Differences in the songs recorded by *Ljungblad* et al. [20.15] and by *Cummings* and *Holliday* [20.149] may be due to bowhead whales changing their songs from year to year.

The sounds from finback whales include single 20 Hz pulses, irregular series of 20 Hz pulses, and stereotyped 20 Hz signal bouts of repetitive sequences of 20 Hz pulses [20.145]. The 20 Hz signals are emitted in bouts that can last for hours. The pulse intervals in a bout were very regular. In general, signals are produced in

Table 20.2 Characteristics of mysticete whales vocalizations

Species of whales	Signal type	Frequency limits(HZ)	Dominat frequency (Hz)	Source level (dB re 1 μPa) at 1 m	References
Blue	FM moans	12.5 – 200	16 – 25	188	*Cummings*, *Thompson* [20.150], *Edds* [20.151]
	Songs	16 – 60	16– 60	–	*McDonald* et al. [20.152]
Bowhead	Tonal moans	25 –900	100 – 400	129 – 178	*Cummings*, *Holliday* [20.149]; *Wursig*, *Clark* [20.148]
	Pulses	25 –3500	152 – 185		
	Songs	20 – 500	158 – 189		*Cummings*, *Holliday* [20.149]; *Ljungblad* et al. [20.15]
Bryde's	FM moans	70 – 245	124 – 132	152 – 174	*Cummings* et al. [20.153]; *Edds* et al. [20.154]
	Pulsed moans	100 - 930	165 – 900	–	*Edds* et al. [20.154]
	Discrete pulses	700- 950	700 – 950	–	*Edds* et al. [20.154]
Finback	FM moans	14 – 118	20	160 – 186	*Watkins* [20.145], *Edds* [20.155], *Cummings*, *Thompson* [20.156]
	Tonals	34 – 150	34 – 150	–	*Edds* [20.155]
	Songs	17 – 25	17 – 25	186	*Watkins* [20.145]
Gray	Pulses	100 – 2000	300 – 825	–	*Dalheim* et al. [20.157]; *Crane* et al. [20.158]
	FM moans	250 – 300	250 – 300	152	*Cummings* et al. [20.159]; *Dalheim* et al. [20.157]
	LF-FM-moans	125 – 1250	< 430	175	*Cummings* et al. [20.159]; *Dalheim* et al. [20.157]
	PM pulses	150 – 1570	225 – 600		*Cummings* et al. [20.159]; *Dalheim* et al. [20.157]
	Complex moans	35 – 360	35 – 360		*Cummings* et al. [20.159]
Humpback	Grunts (pulse & FM)	25 – 1900	25 – 1900	176	*Thompson* et al. [20.160]
	Pulses	25 – 89	25 – 80	144 – 174	*Thompson* et al. [20.160]
	Songs	30 – 8000	120 – 4000	–	*Payne*, *Payne* [20.161]
Minke	FM tones	60 – 130	60 – 130	165	*Schevill*, *Watkins* [20.162],
	Thumps	100 – 200	100 – 200	–	*Winn*, *Perkins* [20.163]
	Grunts	60 – 140	60 – 140	151 – 175	*Winn*, *Perkins* [20.163]
	Ratchets	850 – 6000	850	–	*Winn*, *Perkins* [20.163]
Right-N	Moans	< 400	–	–	*Schevill*, *Watkins* [20.162]
Right-S	Tonal	30 – 1250	160 – 500	–	*Cummings* et al. [20.164], *Clark* [20.129, 165]
	Pulses	30 – 2200	50 – 500	172 – 187	*Cummings* et al. [20.164], *Clark* [20.129, 165]
Sei	FM sweeps	1500 – 3500	1500 – 3500	–	*Knowlton* et al. [20.166]

a relatively regular sequence of repetitions at intervals ranging from about 7–26 s, with bouts that can last as long as 32.5 h. During a bout, periodic rests averaged about 115 s at roughly 15 min intervals and sometimes longer irregular gaps between 20 and 120 min were observed. There was also some variability in the 20 Hz signals in that they were never exactly replicated.

The songs of the blue whales (*balaenoptera musculus*) have been observed by a number of researchers [20.143, 152, 167]. A typical two-part blue whale song time series and corresponding spectrogram is shown in Fig. 20.18. The spectrogram of the first part of the two part song had six spectral lines separated by about 1.5 Hz. This type of spectrogram is typically generated by pulses. *Cummings* and *Thompson* [20.150] previously reported on the pulsive nature of some blue whale moans. The second part of the song was tonal in nature with a slight FM down sweep varying from 19 Hz to 18 Hz in the first 3–4 s. The 18 Hz tone is then carried until the last 5 s when there is an abrupt step down to

Amazon EX1014 Page 835

17 Hz. The amplitude modulation in the second part of the song was probably caused by multi-path propagation of the signal from the whale to the hydrophones. These two-part songs basically follow a pattern of a 19 s pulsive signal followed by a 24.5 s gap and a 19 s monotonic signal [20.152, 167].

Calls of Mysticete Whales

The calls of mysticete whales have been the subject of much research over the past three decades. As with dolphin sounds, there is also a lack of any standard nomenclature for describing emitted sounds. Similar calls are often given different names by different researchers. A summary of some of the acoustic properties of the different baleen whales is indicated in Table 20.2. Calls and songs are most likely used for some sort of communication, however, at this time the specific meanings of these sounds are not known. It is extremely difficult to study the context and functions of baleen whale vocalization. The sounds of mysticete whales have also been summarized nicely by *Richardson* et al. [20.34]. Instead of discussing the characteristics of various sounds, the properties of calls are summarized in Table 20.2. The calls of mysticete whales are mainly in the low-frequency range, from an infrasonic frequency of about 12.5 Hz for the blue whale to about 3.5 kHz for the sei whale. There are a variety of different types of calls, from FM tones to moans, grunts and discrete pulses. How these sounds are used by whales is still an open question since it is often very difficult to observe behavior associated with different calls.

20.5 Discussion

Cetaceans use a wide variety of sounds from brief echolocation signals used by dolphins having durations less than 100 μs to very long duration songs that can last for hours by are emitted by humpback whales. The range of frequency is also very large, from low-frequency (infrasonic) sounds between 10–20 Hz used by blue and finback whales to high-frequency echolocation signals that extend to 130–140 kHz or perhaps higher (see earlier mention of 180 kHz). It is quite clear that acoustics is important in the natural history of cetaceans since all species regularly emit sounds throughout their daily routine. Yet, it is not at all clear how sounds are used by these animals. The difficulty in observing the behavior of these animals has made it difficult to attach any function to a particular or specific sound. It is nearly impossible to determine a one-to-one relationship between the reception or transmission of a specific sound to specific behavioral response. There are many possible functions of sounds, some of which may include providing contact information in a population of animals, signaling alarm or warning of approaching predators, attracting potential mates, echolocation for prey detection and discrimination while foraging, a way of establishing a hierarchy, providing individual identification, and a method for disciplining juveniles. Although researchers have not been successful in uncovering the role or function of specific sounds in any cetacean species, researchers should not be discouraged but should be innovative and imaginative in designing experiments that can be conducted in the wild to delve deeper into this problem. Even with echolocation sounds, there are many unanswered question. It seems logical that echolocation sounds are used to detect and discriminate prey and to navigate, yet we still know very little about how often they are emitted during a daily routine for dolphins in the field.

In this chapter, we have seen that the characteristics of the sounds emitted by cetaceans depends a lot on the size of the animals. Therefore the active space of different species is indirectly related to the size of the animals since the amount of sound absorption by

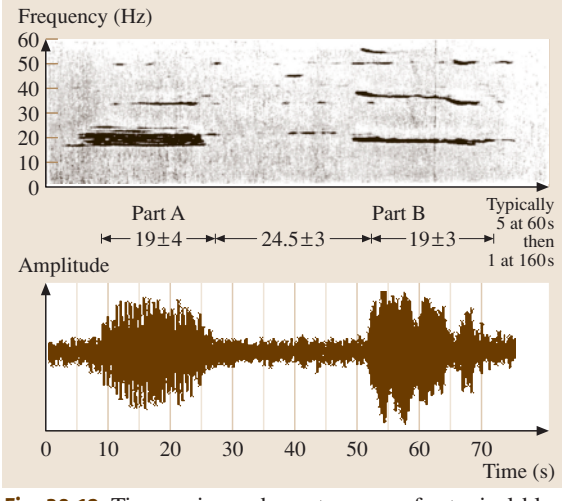


Fig. 20.19 Time series and spectrogram of a typical blue whale song (after [20.152])

sea water increases with frequency. Dolphins and small whales usually emit echolocation signals that can have peak frequencies up to 120–140 kHz and whistles between 10–20 kHz. These signals are not meant to travel large distances so that the active space is rather small, tens of meters for echolocations and burst pulse signals and several hundred meters for whistles. These signals will certainly propagate to much shorter distances than the low-frequency signals used by many of the baleen whales. The sound pressure level (SPL) of an emitted signal at any range from the animal is given by the equation

$$\mathrm{SPL} = \mathrm{SL} - \text{geometric spreading loss} - \alpha R\,, \tag{20.11}$$

where SL is the source level in dB, α is the sound absorption coefficient in dB/m and R is the distance the sound has traveled. The geometric spreading loss does not depend on frequency but is affected by the sound velocity profile, the depth of the source and the receiver. The most severe geometric spreading loss is associated with spherical spreading loss in which the amount of geometric spreading loss is equal to $20 \log R$. The least severe geometric spreading loss is associated with sound channels, such as the sound fixing and ranging (SOFAR) channel. In order to gain an appreciation of the effects of absorption losses for different types of cetacean signals, $-\alpha R$ in (20.9) was calculated as a function of range and the results are shown in Fig. 20.20. The results indicate that the low-frequency sounds used by baleen whales do not suffer as much absorption loss as the sounds of small odontocetes. Therefore, baleen whale signals can propagate long distances if some type of channel is present in the water column.

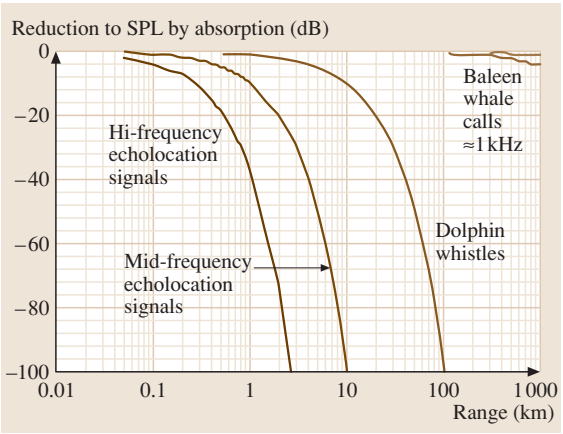


Fig. 20.20 Reduction to sound pressure level caused by sound absorption for different types of cetacean signals. The curve for the high-frequency echolocation signals was calculated at 120 kHz, and for the mid-frequency echolocation signals at 50 kHz. The curve for the dolphin whistle was calculated at 15 kHz. The baleen whale signal curve was determined at 1 kHz.

We have just scraped the top of the iceberg in regards to our understanding of cetacean acoustics. There is much to learn and understand about how these animals produce and receive sounds and how acoustics is used in their daily lives. The future looks extremely promising as new technology arrives on the scene that will allow researchers to delve into the many unanswered questions. As computer chips and satellite transmission tags get smaller and we apply state-of-the-art technologies to the study of cetaceans, our knowledge and understanding can but increase.

References

20.1 S.H. Ridgway: Who are the Whales, Bioacoust. **8**, 3–20 (1997)

20.2 B. Møhl, W.W.L. Au, J. Pawloski, P.E. Nachtigall: Dolphin hearing: Relative sensitivity as a function of point of application of a contact sound source in the jaw and head region, J. Acoust. Soc. Am. **105**, 3421–3424 (1999)

20.3 T.H. Bullock, A.D. Grinnell, E. Ikezono, K. Kameda, Y. Katsuki, M. Nomoto, O. Sato, N. Suga, K. Yanagisawa: Electrophysiological studies of central auditory mechanisms in Cetaceans, Z. Vergl. Phys. **59**, 117–316 (1968)

20.4 J.G. McCormick, E.G. Wever, J. Palin, S.J.H. Ridgway: Sound conduction in the dolphin ear, J. Acoust. Soc. Am. **48**, 1418–1428 (1970)

20.5 K.S. Norris: The evolution of acoustic mechanisms in odontocete cetaceans. In: *Evolution and Environment*, ed. by E.T. Drake (Yale Univ. Press, New Haven 1968) pp. 297–324

20.6 J.L. Aroyan: Three-dimensional modeling of hearing in Delphinus delphis, J. Acoust. Soc. Am. **110**, 3305–3318 (2001)

20.7 W.W.L. Au: Hearing in whales and dolphins: An overview. In: *Hearing by Whales and Dolphins*, ed. by W.W.L. Au, A.N. Popper, R.R. Fay (Springer, Berlin, New York. 2000) pp. 1–42

20.8 W.W.L. Au: Echolocation in dolphins. In: *Hearing by Whales and Dolphins*, ed. by W.W.L. Au, A.N. Popper, R.R. Fay (Springer, Berlin, New York 2000) pp. 364–408

20.9 C.S. Johnson: Sound detection thresholds in marine mammals. In: *Marine Bioacoustics*, ed. by W. Tavolga (Pergamon, New York 1967) pp. 247–260

20.10 S. Andersen: Auditory sensitivity of the harbor porpoise, Phocoena phocoena. Invest. Cetacea **2**, 255–259 (1970)

20.11 R.A. Kastelein, P. Bunskoek, M. Hagedoorn, W.W.L. Au, D. de Haan: Audiogram of a harbor porpoise (Phocoena phocoena) measured with narrow-band frequency modulated signals, J. Acoust. Soc. Am. **112**, 334–344 (2002)

20.12 J.D. Hall, C.S. Johnson: Auditory thresholds of a killer whale, J. Acoust.Soc. Am. **51**, 515–517 (1971)

20.13 M.D. Szymanski, D.E. Bain, K. Kiehi, S. Pennington, S. Wong, K.R. Henry: Killer whale (Orcinus orca) hearing: Auditory brainstem response and behavioral audiograms, J. Acoustic. Soc. Am. **106**, 1134–1141 (1999)

20.14 M.J. White Jr., J. Norris, D. Ljungblad, K. Baron, G. di Sciara: *Auditory thresholds of two beluga whales (Delphinapterus leucas). HSWRI Tech. Rep. No. 78-109* (Hubbs Marine Research Institute, San Diego 1979)

20.15 D.K. Ljungblad, P.D. Scoggins, W.G. Gilmartin: Auditory thresholds of a captive eastern Pacific bottlenose dolphin, Tursiops spp., J. Acoust. Soc. Am. **72**, 1726–1729 (1982)

20.16 J. Thomas, N. Chun, W. Au, K. Pugh: Underwater audiogram of a false killer whale (Pseudorca crassidens), J. Acoust. Soc. Am. **84**, 936–940 (1988)

20.17 D. Wang, K. Wang, Y. Xiao, G. Sheng: Auditory sensitivity of a Chinese River Dolphin, Lipotes vexillifer. In: *Marine Mammal Sensory Systems*, ed. by J.A. Thomas, R.A. Kastelein, Aya Supin (Plenum, New York 1992) pp. 213–221

20.18 P.E. Nachtigall, W.W.L. Au, J.L. Pawloski, P.W.B. Moore: Risso's Dolphin (Grampus griseus) Hearing Thresholds in Kaneohe Bay, Hawaii. In: *Sensory Systems of Aquatic Mammals*, ed. by R.A. Kastelein, J.A. Thomas, P.E. Nachtigall (DeSpiel, Woerden 1995) pp. 49–53

20.19 M. Sauerland, D. Dehnhardt: Underwater Audiogram of a Tucuxi (Sotalia fluviatilis guianensis), J. Acoust. Soc. Am. **103**, 1199–1204 (1998)

20.20 R.A. Kastelein, M. Hagedoorn, W.W.L. Au, D. de Haan: Audiogram of a striped dolphin (Stenella coeruleoalba), J. Acoust. Soc. Am. **113**, 1130–1144 (2003)

20.21 R.R. Fay: *Hearing in Vertebrates: A Psychophysics Databook* (Hill-Fay Associates, Winnetka 1988)

20.22 A.W. Mills: On the Minimum Audible Angle, J. Acoust. Soc. Am. **30**, 237–246 (1958)

20.23 D.L. Renaud, A.N. Popper: Sound localization by the bottlenose porpoise Tursiops truncatus, J. Exp. Biol. **63**, 569–585 (1978)

20.24 A.W. Mills: Auditory localization. In: *Foundations of Modern Auditory Theory*, Vol. 2, ed. by J.V. Tobias (Academic, New York 1972) pp. 303–348

20.25 W.H. van Dudok Heel: Sound and Cetacea, Netherlands J. Sea Res. **1**, 407–507 (1962)

20.26 S. Andersen: Directional hearing in the harbour porpoise, Phocoena phocena. Invest. Cetacea **2**, 261–263 (1970)

20.27 D.L. Renaud, A.N. Popper: Sound localization by the bottlenose porpoise, tursiops truncatus, J. Exp. Biol. **63**, 569–585 (1975)

20.28 W.W.L. Au, P.W.B. Moore: Receiving Beam Patterns and Directivity Indices of the Atlantic Bottlenose Dolphin Tursiops truncates, J. Acoust. Soc. Am. **75**, 255–262 (1984)

20.29 D.S. Houser, J. Finneran, D. Carder, W. Van Bonn, C. Smith, C. Hoh, R. Mattrey, S. Ridgway: Structural and Functional Imaging of BottlenoseDolphin (Tursiops truncatus) Cranial Anatomy, J. Exp. Biol. **207**, 3657–3665 (2004)

20.30 K.A. Zaytseva, A.I. Akopian, V.P. Morozov: Noise resistance of the dolphin auditory analyzer as a function of noise direction, BioFizika **20**, 519–521 (1975)

20.31 R.J. Urick: *Principles of Underwater Sounds for Engineers* (McGraw-Hill, New York 1980)

20.32 J.M. McCormick, M.G. Salvadori: *Numerical Methods in FORTRAN* (Prentice-Hall, Englewood Cliffs 1964),

20.33 D.R. Ketten: The marine mammal ear: specializations for aquatic audition and echolocation. In: *The Evolutionary Biology of Hearing*, ed. by D. Webster, R. Fay, A. Popper (Springer, Berlin, New York 1992) pp. 717–754

20.34 W.J. Richardson, C.R. Greene Jr., C.I. Malme, D.H. Thomson: *Marine Mammals and Noise* (Academic, San Diego 1995)

20.35 W.J. Richardson, B. Würsig, C.R. Greene Jr: Reactions of bowhead whales, Balaena mysticetus, to drilling and dredging noise in the Canadian Beaufort Sea, Mar. Envr. Res. **29**, 135–160 (1990),

20.36 A.S. Frankel: *Acoustic and visual tracking reveals distribution, song variability and social roles of humpback whales in Hawai'ian waters. Ph.D. Dissertation* (Univ. Hawaii, Honolulu, HI 1994)

20.37 C.R. Ridgway: Electrophysiological experiments on hearing in odontocetes. In: *Animal Sonar*, ed. by P.E. Nachtigall, P.W.B. Moore (Plenum, New York 1980) pp. 483–493

20.38 S.H. Ridgway, D.A. Carder: Assessing hearing and sound production in cetaceans not available for behavioral audiograms: Experiences with sperm, pygmy sperm and gray whales, Aquatic Mammals. **27**, 267–276 (2001)

20.39 W.W.L. Au: *The Sonar of Dolphins* (Springer, Berlin, New York 1993)

20.40 S.H. Ridgway, D.A. Carder: Nasal pressure and sound production in an echolocating white whale, Delphinapterus leucas. In: *Animal Sonar: Processes and Performance*, ed. by P.E. Nachtigall,

Amazon EX1014 Page 838

P.W.B. Moore (Plenum, New York 1988) pp. 53–60,
20.41 F.G. Wood: Discussion. In: *Marine Bio-Acoustics Vol. II*, ed. by W. Tavolga (Pergamon, Oxford 1964) pp. 395–396
20.42 K.S. Norris, G.W. Harvey: Sound transmission in the porpoise head, J. Acoust. Soc. Am. **56**, 659–664 (1974)
20.43 J.L. Aroyan, T.W. Cranford, J. Kent, K.S. Norris: Computer modeling of acoustic beam formation in Delphinus delphis, J. Acoust. Soc. Am. **95**, 2539–2545 (1992)
20.44 T.W. Cranford: In Search of Impulse sound sources in odontocetes. In: *Hearing by Whales and Dolphins*, ed. by W.W.L. Au, A.N. Popper, R.R. Fay (Springer, New York 2000) pp. 109–155
20.45 W.W.L. Au, R.W. Floyd, R.H. Penner, A.E. Murchison: Measurement of echolocation signals of the Atlantic bottlenose dolphin Tursiops truncatus Montagu, in Open Waters, J. Acoust. Soc. Am. **56**, 280–1290 (1974)
20.46 W.E. Evans: Echolocation by marine delphinids and one species of fresh-water dolphin, J. Acoust. Soc. Am. **54**, 191–199 (1973)
20.47 W.W.L. Au, D.A. Carder, R.H. Penner, B.L. Scronce: Demonstration of adaptation in beluga whale echolocation signals, J. Acoust. Soc. Am. **77**, 726–730 (1985)
20.48 B.S. Gurevich, W.E. Evans: Echolocation discrimination of complex planar targets by the beluga whale (Delphinapterus leucas), J. Acoust. Soc. Am. **60**, S5 (1976)
20.49 C. Kamminga, H. Wiersma: Investigations on cetacean sonar V. The true nature of the sonar sound of Cephaloryncus Commersonii, Aqua. Mamm. **9**, 95–104 (1981)
20.50 C.W. Turl, D.J. Skaar, W.W.L. Au: The Echolocation Ability of the Beluga (Delphinapterus leucas) to detect Targets in Clutter, J. Acoust. Soc. Am. **89**, 896–901 (1991)
20.51 P.W.B. Moore, D.A. Pawloski: Investigations on the control of echolocation pulses in the dolphin (Tursiops truncatus). In: *Sensory Abilities of Cetaceans Laboratory and Field Evidence*, ed. by J.A. Thomas, R.A. Kastelein (Plenum, New York 1990) pp. 305–316
20.52 W.W.L. Au, J.L. Pawloski, P.E. Nachtigall: Echolocation signals and transmission beam pattern of a false killer whale (Pseudorc crassidens), J. Acoust. Soc. Am. **98**, 51–59 (1995)
20.53 J. Thomas, M. Stoermer, C. Bowers, L. Anderson, A. Garver: Detection abilities and signal characteristics of echolocating false killer whales (Pseudorca crassidens). In: *Animal Sonar Processing and Performance*, ed. by P.E. Nachtigall, P.W.B. Moore (Plenum, New York 1988) pp. 323–328
20.54 J.A. Thomas, C.W. Turl: Echolocation characteristics and range detection by a false killer whale Pseudorca crassidens. In: *Sensory Abilities of Cateceans: Laboratory and Field Evidence*, ed. by J.A. Thomas, R.A. Kastelein (Plenum, New York 1990) pp. 321–334
20.55 W.W.L. Au, J.K.B. Ford, J.K. Horne, K.A. Newman-Allman: Echolocation signals of free-ranging killer whales (Orcinus orca) and modeling of foraging for chinook salmon (Oncorhynchus tshawytscha), J. Acoust. Soc. Am. **115**, 901–909 (2004)
20.56 W.W.L. Au, B. Würsig: Echolocation signals of dusky dolphins (Lagenorhynchus obscurus) in Kaikoura, New Zealand, J. Acoust. Soc. Am. **115**, 2307–1313 (2004)
20.57 W.W.L. Au, D. Herzing: Echolocation signals of wild Atlantic spotted dolphin (Stenella frontalis), J. Acoust. Soc. Am. **113**, 598–604 (2003)
20.58 W.A. Watkins, W.E. Schevill: Sperm whale codas, J. Acoust. Soc. Am. **62**, 1485–1490 (1977)
20.59 C. Levenson: Source level and bistatic target strength of sperm whale (Physeter catodon) measured from an oceanographic aircraft, J. Acoust. Soc. Am. **55**, 1100–1103 (1974)
20.60 L. Weilgart, H. Whitehead: Distinctive vocalizations from mature male sperm whales (Physter macrocephalus), Can. J. Zool. **66**, 1931–1937 (1988)
20.61 B. Møhl, E. Larsen, M. Amundin: Sperm Whale Size Determination: Outlines of an Acoustic Approach. In: *Mammals in the Seas. Fisheries series No. 5, Mammals of the Seas*, Vol. 3 (FAO (Food and Agriculture Organization of the United Nations) Publications, Rome 1981) pp. 327–331
20.62 J.L. Dunn: Airborne Measurements of the Acoustic Characteristics of a Sperm Whale, J. Acoust. Soc. Am. **46**, 1052–1054 (1969)
20.63 W.A. Watkins: Acoustics and the behavior of sperm whales. In: *Animal Sonar Systems*, ed. by R.G. Busnel, J.F. Fish (Plenum, New York 1980) pp. 291–297
20.64 B. Møhl, M. Wahlberg, A. Heerfordt: A GPS-linked array of independent receiver for bioacoustics, J. Acoust. Am. **109**, 434–437 (2001)
20.65 B. Møhl, M. Wahlberg, P.T. Madsen, A. Heerfordt, A. Lund: The monpulsed natue of sperm whale clicks, J. Acoust. Soc. Am. **114**, 1143–1154 (2003)
20.66 W.W.L. Au, R.H. Penner, C.W. Turl: Propagation of beluga echolocation signal, J. Acoust. Soc. Am. **82**, 807–813 (1987)
20.67 R.A. Kastelein, W.W.L. Au, H.T. Rippe, N.M. Schooneman: Target detection by an echolocating harbor porpoise (Phocoena phocoena), J. Acoustic. Soc. Am. **105**, 2493–2498 (1999)
20.68 W.W.L. Au, K.J. Snyder: Long-range Target Detection in Open Waters by an Echolocating Atlantic Bottlenose Dolphin (Tursiops truncatus), J. Acoust. Soc. Am. **68**, 1077–1084 (1980)
20.69 P.T. Madsen, D.A. Carder, K. Bedholm, S.H. Ridgway: Porpoise clicks from a sperm whale nose –convergent evolution of 130 kHz pulses in toothed whale sonars?, Bioacoustics **15**, 192–206 (2005)
20.70 W.W.L. Au, R.A. Kastelein, T. Ripper, N.M. Schooneman: Transmission beam pattern and echolocation

Amazon EX1014 Page 839

signals of a harbor porpoise (Phocoena phocoena), J. Acoustic. Soc. Am. **106**, 3699–3705 (1999)
20.71 K.W. Pryor, K.S. Norris: *Dolphin Societies: Discoveries and Puzzles* (Univ. California Press, Berkley 1991)
20.72 K.S. Norris, B. Würsig, R.S. Wells, M. Würsig: *The Hawai'ian Spinner Dolphin* (Univ. California Press, Berkeley 1994) p. 408
20.73 J. Mann, P.L. Tyack, H. Whitehead: *Cetacean societies: field studies of dolphins and whales* (Univ. Chicago Press, Chicago 2000)
20.74 L.M. Herman, W.N. Tavolga: The communications systems of cetaceans. In: *Cetacean Behavior: Mechanisms and Function*, ed. by L.M. Herman (Wiley-Interscience, New York 1980) pp. 149–209
20.75 K.W. Pryor: Non-acoustic communication in small cetaceans: glance touch position gesture and bubbles. In: *Sensory Abilities of Cetaceans*, ed. by J. Thomas, R. Kastelein (Plenum, New York 1990) pp. 537–544
20.76 K.M. Dudzinski: Contact behavior and signal exchange in Atlantic spotted dolphins (Stenella frontalis), Aqua. Mamm. **24**, 129–142 (1998)
20.77 C.M. Johnson, K. Moewe: Pectoral fin preference during contact in Commerson's dolphins (Cephalorhynchus commersonii), Aqua. Mamm. **25**, 73–77 (1999)
20.78 P.E. Nachtigall: Vision, audition and chemoreception in dolphins and other marine mammals. In: *Dolphin Cognition and Behavior: A Comparative Approach*, ed. by R.J. Schusterman, J.A. Thomas, F.G. Wood (Lawrence Erlbaum, Hillsdale 1986) pp. 79–113
20.79 V.B. Kuznetsov: Chemical sense of dolphins: quasi-olfaction. In: *Sensory Abilities of Cetaceans*, ed. by J. Thomas, R. Kastelein (Plenum, New York 1990) pp. 481–503
20.80 K.S. Norris, E.C. Evans III.: On the evolution of acoustic communication systems in vertebrates, Part I: Historical aspects. In: *Animal Sonar: Processes and Performance*, ed. by P.E. Nachtigall, P.W.B. Moore (Plenum, New York 1988) pp. 655–669
20.81 R.E. Thomas, K.M. Fristrup, P.L. Tyack: Linking the sounds of dolphins to their locations and behavior using video and multichannel acoustic recordings, J. Acoust. Soc. Am. **112**, 1692–1701 (2002)
20.82 K.R. Ball, J.R. Buck: A beamforming video recorder for integrated observations of dolphin behavior and vocalizations, J. Acoust. Soc. Am. **117**, 1005–1008 (2005)
20.83 M.O. Lammers, W.W.L. Au, D.L. Herzing: The broadband social acoustic signaling behavior of spinner and spotted dolphins, J. Acoust. Soc. Am. **114**, 1629–1639 (2003)
20.84 M.O. Lammers, W.W.L. Au: Directionality in the whistles of Hawai'ian spinner dolphins (Stenella longirostris): A signal feature to cue direction of movement?, Mar. Mamm. Sci. **19**, 249–264 (2003)
20.85 V.M. Janik, P.J.B. Slater: Context specific use suggests that bottlenose dolphin signature whistles are cohesion calls, Anim. Behav. **56**, 829–838 (1998)
20.86 W.W. Steiner: Species-specific differences in pure-tonal whistle vocalizations of five western north Atlantic dolphin species, Behav. Ecol. Sociobiol. **9**, 241–246 (1981)
20.87 L.E. Rendell, J.N. Matthews, A. Gill, J.C.D. Gordon, D.W. MacDonald: Quantitative analysis of tonal calls from five odontocete species, examining interspecific and intraspecific variation, J. Zool. **249**, 403–410 (1999)
20.88 W. Ding, B. Wursig, W.E. Evans: Comparisons of whistles among seven odontocete species. In: *Sensory Systems of Aquatic Mammals*, ed. by R.A. Kastelein, J.A. Thomas (De Spil, Woerden 1995) pp. 299–323
20.89 W. Ding, B. Wursig, W.E. Evans: Whistles of bottlenose dolphins: comparisons among populations, Aqua. Mamm. **21**, 65–77 (1995)
20.90 C. Bazua-Duran, W.W.L. Au: Geographic variations in the whistles of spinner dolphins (Stenella longirostris) of the Main Hawai'ian Islands, J. Acoust. Soc. Am. **116**, 3757–3769 (2004)
20.91 J.N. Oswald, J. Barlow, T.F. Norris: Acoustic identification of nine delphinid species in the eastern tropical Pacific Ocean, Mar. Mamm. Sci. **19**, 20–37 (2003)
20.92 P.J.O. Miller: Mixed-directionality of killer whale stereotyped calls: a direction-of-movement cue?, Behav. Ecol. Sociobiol. **52**, 262–270 (2002)
20.93 M.O. Lammers, W.W.L. Au, R. Aubauer: A comparative analysis of echolocation and burst-pulse click trains in Stenella longirostris. In: *Echolocation in Bats and Dolphins*, ed. by J. Thomas, C. Moss, M. Vater (Univ. Chicago Press, Chicago 2004) pp. 414–419
20.94 A.E. Puente, D.A. Dewsbury: Courtship and copulatory behavior of bottlenose dolphins (Tursiops truncatus), Cetology **21**, 1–9 (1976)
20.95 M.C. Caldwell, D.K. Caldwell: Intraspecific transfer of information via the pulsed sound in captive Odontocete Cetaceans. In: *Animal Sonar Systems: Biology and Bionics*, ed. by R.G. Busnel (Laboratoire de Physiologie Acoustic: Jouy-en-Josas, France 1967) pp. 879–936
20.96 S.M. Dawson: Clicks and communication: the behavioural and social contexts of Hector's dolphin vocalizations, Ethology **88**, 265–276 (1991)
20.97 D.L. Herzing: Vocalizations and associated underwater behavior of free-ranging Atlantic spotted dolphins, Stenella frontalis and bottlenose dolphin, Tursiops truncatus, Aqua. Mamm. **22**, 61–79 (1996)

Amazon EX1014 Page 840

20.98 M.C. Caldwell, D.K. Caldwell: Underwater pulsed sounds produced by captive spotted dolphins, Stenella plagiodon, Cetology **1**, 1–7 (1971)

20.99 J.K.B. Ford, H.D. Fisher: Underwater acoustic signals of the narwhal (Monodon monoceros), Can. J. Zool. **56**, 552–560 (1978)

20.100 S.M. Dawson: The high frequency sounds of free-ranging hector's dolphins, Cephalorhynchus hectori, Rep. Int. Whal. Comm. **Spec. Issue 9**, 339–344 (1988)

20.101 R.G. Busnel, A. Dziedzic: Acoustic signals of the Pilot whale Globicephala melaena, Delpinus delphis and Phocoena phocoena. In: *Whales, Dolphins and porpoises*, ed. by K.S. Norris (Univ.California Press, Berkley 1966) pp. 607–648

20.102 N.A. Overstrom: Association between burst-pulse sounds and aggressive behavior in captive Atlantic bottlenose dolphins (Tursiops truncatus), Zoo Biol. **2**, 93–103 (1983)

20.103 B. McCowan, D. Reiss: Maternal aggressive contact vocalizations in captive bottlenose dolphins (Tursiops truncatus): wide-band, low frequency signals during mother/aunt-infant interactions, Zoo Biol. **14**, 293–309 (1995)

20.104 C. Bloomqvist, M. Amundin: High frequency burst-pulse sounds in agonistic/ aggressive interactions in bottlenose dolphins (Tursiops truncatus). In: *Echolocation in Bats and Dolphins*, ed. by J. Thomas, C. Moss, M. Vater (Univ.Chicago Press, Chicago, IL 2004) pp. 425–431

20.105 J.C. Lilly, A.M. Miller: Sounds emitted by the bottlenose dolphin, Science **133**, 1689–1693 (1961),

20.106 D.L. Herzing: *A Quantitative description and behavioral association of a burst-pulsed sound, the squawk, in captive bottlenose dolphins, Tursiops truncatus, Masters Thesis* (San Francisco State University, San Francisco 1988) p. 87

20.107 S.M. Brownlee: *Correlations between sounds and behavior in wild Hawai'ian spinner dolphins (Stenella longirostris). Masters Thesis* (Univ. California, Santa Cruz 1983) p. 26

20.108 V.A. Vel'min, N.A. Dubrovskiy: The critical interval of active hearing in dolphins, Sov. Phys. Acoust. **2**, 351–352 (1976)

20.109 V.M. Janik: Source levels and the estimated active space of bottlenose dolphin (Tursiops truncatus) whistles in the Moray Firth, Scotland, J. Comp. Psych. **186**, 673–680 (2000)

20.110 S.O. Murray, E. Mercado, H.L. Roitblat: Characterizing the graded structure of false killer whale (Pseudorca crassidens) vocalizations, J. Acoust. Soc. Am. **104**, 1679–1688 (1998)

20.111 M.C. Caldwell, D.K. Caldwell: Vocalizations of naïve captive dolphins in small groups, Science **159**, 1121–1123 (1968)

20.112 W.A. Powell: Periodicity of vocal activity of captive Atlantic bottlenose dolphins: Tursiops truncatus, Bull. Soc. Ca. Acad. Sci. **65**, 237–244 (1966),

20.113 I.E. Sidorova, V.I. Markov, V.M. Ostrovskaya: Signalization of the bottlenose dolphin during the adaptation to different stressors. In: *Sensory Abilities of Cetaceans*, ed. by J. Thomas, R. Kastelein (Plenum, New York 1990) pp. 623–638

20.114 S.E. Moore, S.H. Ridgway: Whistles produced by common dolphins from the Southern California Bight, Aqua. Mamm. **21**, 55–63 (1995)

20.115 J.C. Goold: A diel pattern in vocal activity of short-beaked common dolphins, Delphinus delphis, Mar. Mamm. Sci. **16**, 240–244 (2000)

20.116 M.C. Caldwell, D.K. Caldwell: Individualized whistle contours in bottlenosed dolphins (Tursiops truncatus), Nature **207**, 434–435 (1965)

20.117 P.L. Tyack: Whistle repertoires of two bottlenose dolphins, Tursiops truncatus: mimicricy of signature whistles?, Behav. Ecol. Sociobiol. **18**, 251–257 (1986)

20.118 M.C. Caldwell, D.K. Caldwell, P.L. Tyack: A review of the signature whistle hypothesis for the Atlantic bottlenose dolphin. In: *The Bottlenose Dolphin*, ed. by S. Leatherwood, R.R. Reeves (Academic, San Diego 1990) pp. 199–234

20.119 L.S. Sayigh, P.L. Tyack, R.S. Wells, M.D. Scott, A.B. Irvine: Signature differences in signature whistles production of free-ranging bottlenose dolphins Tursiops truncatus, Behav. Ecol. Sociobiol. **36**, 171–177 (1990)

20.120 L.S. Sayigh, P.L. Tyack, R.W. Wells, M.D. Scott, A.B. Irvine: Sex difference in signature whistle production of free-ranging bottlenose dolphins, Tursiops truncatus, Behav. Ecol. Sociobiol. **36**, 171–177 (1995)

20.121 V.M. Janik, G. Denhardt, T. Dietmar: Signature whistle variation in a bottlenosed dolphin, Tursiops truncatus, Behav. Ecol. Sociobiol. **35**, 243–248 (1994)

20.122 R.A. Smolker, J. Mann, B.B. Smuts: Use of signature whistles during separations and reunions by wild bottlenose dolphin mothers and infants, Behav. Ecol. Sociobiol. **33**, 393–402 (1993)

20.123 V.M. Janik: Whistle matching in wild bottlenose dolphins (Tursiops truncatus), Science **289**, 1355–1357 (2000)

20.124 B. McCowan, D. Reiss: Quantitative comparison of whistle repertoires from captive adult bottlenose dolphins (Delphinidae, Tursiops truncatus): A re-evaluation of the signature whistle hypothesis, Ethology **100**, 194–209 (1995)

20.125 B. McCowan, D. Reiss: The fallacy of 'signature whistles' in bottlenose dolphins: a comparative perspective of 'signature information' in animal vocalizations, Anim. Behav. **62**, 1151–1162 (2001)

20.126 J.W. Bradbury, S.L. Vehrencamp: *Principles of Animal Communication* (Sinauer Associates, Sunderland 1998) p. 882

Amazon EX1014 Page 841

20.127 D. Reiss, B. McCowan: Spontaneous vocal mimicry and production by bottlenose dolphins (Tursiops truncatus): Evidence for vocal learning, J. Comp. Psychol. **101**, 301–312 (1993)

20.128 A.G. Taruski: The whistle repertoire of the North Atlantic pilot whale (Globicephala melaena) and its relationship to behavior and environment. In: *Behavior of Marine Animals: Current Perspectives on Research. Vol. 3: Cetaceans*, ed. by H.E. Winn, B.L. Olla (Plenum, New York 1979) pp. 345–368

20.129 C.W. Clark: The acoustic repertoire of the southern right whale: A quantitative analysis, Anim. Behav. **30**, 1060–1071 (1982)

20.130 D.G. Richards, J.P. Wolz, L.M. Herman: Vocal mimicry of computer-generated sounds and vocal labeling of objects by a bottlenose dolphin, Tursiops truncatus, J. Comp. Psych. **98**, 10–28 (1984)

20.131 V.B. Deecke, J.K.B. Ford, P. Spong: Dialect change in resident killer whales: Implications for vocal learning and cultural transmission, Anim. Behav. **60**, 629–638 (2000)

20.132 B. McCowan, D. Reiss, C. Gubbins: Social familiarity influences whistle acoustic structure in adult female bottlenose dolphins (Tursiops truncatus), Aqua. Mamm. **24**, 27–40 (1998)

20.133 L.G. Grimes: Dialects and geographical variation in the song on the splendid sunbird Nectarinia coccinigaster, Ibis **116**, 314–329 (1974)

20.134 J.K.B. Ford: Vocal traditions among resident killer whales (Orcinus orca) in coastal waters of British Columbia, Can. J. Zool. **69**, 1454–1483 (1991)

20.135 F. Nottebohm: The song of the chingolo, Zonotrichia capensis, in Agentina: description and evaluation of a system of dialect, Condor **71**, 299–315 (1969)

20.136 J.K.B. Ford: Acoustic behaviour of resident killer whales (Orcinus orca) off Vancouver Island, British.Columbia, Can. J. Zool. **67**, 727–745 (1989)

20.137 J.K.B. Ford, H.D. Fisher: Group-specific dialects of killer whales (Orcinus orca) in British Columbia. In: *Communication and Behavior of Whales. AAAS Sel Symp. 76. Boulder* (Westview Press, Boulder, CO 1983) pp. 129–161

20.138 J.K.B. Ford: *Call traditions and dialects of killer whales (Orcinus orca) in British Columbia, Ph.D. Dissertation* (Univ. British Columbia, Vancouver, BC 1984)

20.139 A.E. Bowles, W.G. Young, E.D. Asper: On ontogeny of stereotyped calling of a killer whale calf, orcinus orca, during her first year, Rit Fiskideildar **11**, 251–275 (1988)

20.140 H. Whitehead, S. Waters, L. Weilgart: Social organization in female sperm whales nd their offspring: Constant companions and casual acquaintance, Behav. Ecol. Sociobiol. **29**, 385–389 (1991)

20.141 L. Weilgart, H. Whitehead: Group-Specific Dialects and Geographical Variation in Coda Repertoire in South Pacific Sperm Whale, Behav. Ecol. Sociobiol. **40**, 277–285 (1997)

20.142 C.W. Clark: Acoustic behavior of mysticete whales. In: *Sensory Abilities of Cetaceans*, ed. by J. Thomas, R. Kastelein (Plenum, New York 1990) pp. 571–583

20.143 A.E. Alling, E.M. Dorsey, J.C. Gordon: Blue whales (Balaenoptera musculus) off the northeast coast of Sri Lanka: distribution, feeding, and individual identification. In: *Cetaceans and Cetacean Research in the Indian Ocean Sanctuary*, Mar. Mammal Tech. Rep., Vol. 3, ed. by S. Leaderwood, G.P. Donovan (UNEP, New York, NY 1991) pp. 247–258

20.144 W.A. Watkins, P.L. Tyack, K.E. Moore, J.E. Bird: The 20-Hz signals of finback whales (Balaenoptera physalus), J. Acoust. Soc. Am. **82**, 1901–1912 (1987)

20.145 W.A. Watkins: The activities and underwater sounds of fin whales, Sci. Rep. Whales Res. Inst. **33**, 83–117 (1981)

20.146 R.S. Payne, S. Mcvay: Songs of humpback whales. Humpbacks emit sounds in long, predictable patterns ranging over frequencies audible to humans, Science **173**, 585–597 (1971)

20.147 D.A. Helweg, A.S. Frankel, J.R. Mobley Jr, L.M. Herman: Humpback whale song: our current understanding. In: *Sensory Abilities of Aquatic Mammals*, ed. by J.A. Thomas, R.A. Kastelein, Aya Supin (Plenum, New York 1992) pp. 459–483

20.148 B. Wursig, C. Clark: Behavior. In: *The Bowhead Whale*, Lawrence Soc. Mar. Mammal Spec. Publ., Vol. 2, ed. by J.J. Burns, J.J. Montague, C.J. Cowles (Allen, Lawrence 1993) pp. 157–199

20.149 W.C. Cummings, D.V. Holliday: Sounds and source levels from bowhead whales off Pt. Barrow, Alaska. J. Acoust. Soc. Am. **82**, 814–821 (1987)

20.150 W.C. Cummings, P.O. Thompson: Underwater Sounds from the Blue Whale, Balaenoptera musculus, J. Acoust. Soc. Am. **50**, 1193–1198 (1971)

20.151 P.L. Edds: Characteristics of the blue whale, balaenoptera musculus, in St. Lawrence river, J. Mamm. **63**, 345–347 (1982)

20.152 M.A. McDonald, J.A. Hilderbrand, S.C. Webb: Blue and fin whales observed on a seafloor array in the northeast Pacific, J. Acoust. Soc. Am. **98**, 712–721 (1995)

20.153 W.C. Cummings, P.O. Thompson, S.J. Ha: Sounds from bryde's whale balaenoptera edeni, and finback, b. physalus, whales in the gulf of California: Fishery bulletin, Breviora Mus. Comp. Zool. **84**, 359–370 (1986)

20.154 P.L. Edds, D.K. Odell, B.R. Tershy: Vocalization of a captive juvenile and free-ranging adult-calf pairs of bryde's whales, balaenoptera edeni, Mar. Mamm. Sci. **9**, 269–284 (1993)

20.155 P.L. Edds: Characteristic of finback, balaenoptera physalus, vocalization in the St. Lawrence estuary, Bioacoustics **1**, 131–149 (1988)

Amazon EX1014 Page 842

20.156 W.C. Cummings, P.O. Thompson: Characteristics and seasons of blue and finback whale sounds along the U.S. west coast as recorded at SOSUS stations, J. Acoust. Soc. Am. **95**, 2853 (1994)
20.157 M.E. Dalheim, D.H. Fisher, J.D. Schempp: Sound Production by the Grey Whale and Ambient Noise Levels in Laguna San Ignacio, Baja California Sur, Mexico. In: *The Grey Whale*, ed. by M. Jones, S. Swartz, S. Leatherwood (Academic Press, S.D. 1984) pp. 511–541
20.158 N.L. Crane: *Sound Production of Gray Whales, Eschrichtius tobustus Along Their Migration Route* (MS Thesis, San Francisco St. Univ., San Francisco, CA 1992)
20.159 W.C. Cummings, P.O. Thompson, R. Cook: Underwater sounds of migrating gray whales, eschrichtius glaucus (cope), J. Acoust. Soc. Am. **44**, 278–1281 (1968)
20.160 P.O. Thompson, W.C. Cummings, S.J. Ha: Sounds, Source Levels, and Associated Behavior of Humpback Whales, Southeast Alaska, J. Acoust. Soc. Am. **80**, 735–740 (1986)
20.161 R.S. Payne, K. Payne: Large scale changes over 19 years in songs of humpback whales in bermuda, Z. Tierpsychol. **68**, 89–114 (1985)
20.162 W.E. Schevill, W.A. Watkins: Intense low-frequency sounds from an antarctic minke whale, balaenoptera acutorostrata, Breviora Mus. Comp. Zool., 1–8 (1972)
20.163 H.E. Winn, P.J. Perkins: Distribution and sounds of the minke whale, with a review of mysticete sounds, Cetology **19**, 1–11 (1976)
20.164 W.C. Cummings, J.F. Fish, P.O. Thompson: Sound production and other behavior of southern right wales, eubalena glacialis, Trans. San Diego Soc. Nat. Hist. **17**, 1–13 (1972)
20.165 C.W. Clark: Acoustic Communication and Behavior of the Southern Right Whale. In: *Behavior and Communication of Whales*, ed. by R.S. Payne (Westview Press, Boulder, CO 1983) pp. 643–653
20.166 A. R. Knowlton, W. W. Clark, S. D. Kraus: Sounds Recorded in the Presence of Sei Whales, Balaenoptera borealis, in Proc. 9th Bienn. Conf. Biol. Mar. Mamm., Chicago, IL, Dec. 1991, p. 40 (1991)
20.167 P.O. Thompson, L.T. Findlay, O. Vidal: Underwater sounds of blue whales, Balaenoptera musculus, in the Gulf of California, Mexico, Mar. Mamm. Sci. **12**, 288–293 (1987)

Amazon EX1014 Page 843

21. Medical Acoustics

Medical acoustics can be subdivided into diagnostics and therapy. Diagnostics are further separated into auditory and ultrasonic methods, and both employ low amplitudes. Therapy (excluding medical advice) uses ultrasound for heating, cooking, permeablizing, activating and fracturing tissues and structures within the body, usually at much higher amplitudes than in diagnostics. Because ultrasound is a wave, linear wave physics are generally applicable, but recently nonlinear effects have become more important, even in low-intensity diagnostic applications.

This document is designed to provide the nonmedical acoustic scientist or engineer with some insights into acoustic practices in medicine. Auscultation with a stethoscope is the most basic use of acoustics in medicine and is dependent on the fields of incompressible (circulation) and compressible (respiration) fluid mechanics and frictional mechanics. Detailed discussions of tribology, laminar and turbulent hemodynamics, subsonic and supersonic compressional flow, and surfactants and inflation dynamics are beyond the scope of this document. However, some of the basic concepts of auscultation are presented as a starting point for the study of natural body sounds. Ultrasonic engineers have dedicated over half a century of effort to the development of ultrasound beam patterns and beam scanning methods, stretching the current technical and economic limits of analog and digital electronics and signal processing at each stage. The depth of these efforts cannot be covered in these few pages. However, the basic progression of progress in the fields of transducers and signal processing will be covered. The study of the interaction of ultrasound with living tissues is complicated by complex anatomic structures, the high density of scatterers, and the constantly changing nature of the tissues with ongoing life processes including cardiac pulsations, the formation of edema and intrinsic noise sources. A great deal of work remains to be done on the ultrasonic characterization of tissues. Finally, the effect of ultrasound on tissues, both inadvertent and therapeutic will be discussed.

Much of the medical acoustic literature published since 1987 is searchable online, so this document has included key words that will be helpful in performing a search. However, much of the important basic work was done before 1987. In an attempt to help the reader to access that literature, Denis White and associates have compiled a complete bibliography of the medical ultrasound literature prior to 1987. Under Further Reading in this chapter, the reader will find a link to a complete compilation of 99 citations from *Ultrasound in Medicine and Biology* which list the thousands of articles on medical acoustics written prior to 1987.

The academically based authors develop, use and commercialize diagnostic ultrasonic Doppler systems for the benefit of patients with cardiovascular diseases. To translate ultrasonic and acoustic innovation into widespread clinical application requires as much knowledge about the economics of medicine, the training and practices of medical personnel, and the pathology and prevalence of diseases as about the diffraction patterns of ultrasound beams and signal-to-noise ratio of an echo. Although a discussion of these factors is beyond the scope of this chapter, a few comments will help to provide perspective on the likely future contribution of medical acoustics to improved public health.

21.1 **Introduction to Medical Acoustics** 841

21.2 **Medical Diagnosis; Physical Examination** 842
- 21.2.1 Auscultation – Listening for Sounds 842
- 21.2.2 Phonation and Auscultation 847
- 21.2.3 Percussion 847

21.3 **Basic Physics of Ultrasound Propagation in Tissue** 848
- 21.3.1 Reflection of Normal-Angle-Incident Ultrasound 850

Amazon EX1014 Page 844

21.3.2 Acute-Angle Reflection of Ultrasound 850
21.3.3 Diagnostic Ultrasound Propagation in Tissue 851
21.3.4 Amplitude of Ultrasound Echoes . 851
21.3.5 Fresnel Zone (Near Field), Transition Zone, and Fraunhofer Zone (Far Field) .. 853
21.3.6 Measurement of Ultrasound Wavelength 855
21.3.7 Attenuation of Ultrasound 855

21.4 **Methods of Medical Ultrasound Examination** 857
21.4.1 Continuous-Wave Doppler Systems 857
21.4.2 Pulse-Echo Backscatter Systems .. 859
21.4.3 B-mode Imaging Instruments 862

21.5 **Medical Contrast Agents** 882
21.5.1 Ultrasound Contrast Agents 883
21.5.2 Stability of Large Bubbles 883
21.5.3 Agitated Saline and Patent Foramen Ovale (PFO) . 884
21.5.4 Ultrasound Contrast Agent Motivation 886
21.5.5 Ultrasound Contrast Agent Development 886
21.5.6 Interactions Between Ultrasound and Microbubbles 886
21.5.7 Bubble Destruction 887

21.6 **Ultrasound Hyperthermia in Physical Therapy** 889

21.7 **High-Intensity Focused Ultrasound (HIFU) in Surgery** 890

21.8 **Lithotripsy of Kidney Stones** 891

21.9 **Thrombolysis** 892

21.10 **Lower-Frequency Therapies** 892

21.11 **Ultrasound Safety** 892

References 895

Standard Symbols

c sound propagation velocity [km/s = mm/μs]
p acoustic pressure fluctuation [d/cm^2; Pa = N/m^2]
t time [s]
Z tissue impedance [Rayles = ρc = kg/(m^{2}s) = 0.1 g/(cm^{2}s)]
λ acoustic wavelength [m]
ρ tissue density [g/cm^3]
$\boldsymbol{\sigma}$ surface tension [d/cm = mN/m]

Special Symbols

d diameter
f frequency [Hz]
f_D Doppler frequency [Hz]
f_{us} ultrasound center frequency [MHz]
fd vibration frequency [Hz]
P total instantaneous pressure [d/cm^2, kPa]
q acoustic molecular displacement [cm]
Q volume flow [cm^3/s]
R reflection coefficient
T transmission coefficient
v_f fluid velocity [cm/s]
v molecular velocity [cm/s]
w transducer element width [cm]
x distance in propagation direction [cm]
κ tissue stiffness [d/cm^2]
θ angle of sound propagation from the normal at a refraction interface
Ψ distance to transition zone
θ_D Doppler angle of sound propagation to vessel axis or velocity vector axis
$\Delta\phi$ ultrasound echo phase shift
$\Delta\boldsymbol{\tau}$ ultrasound echo time shift [ns]
PRF pulse repetition frequency [Hz]
$\boldsymbol{D}$ thermal dose

Subscripts

i incident
r reflected
s shear wave
t transmitted
1 incident material
2 second material

Dimensionless Numbers

St Strouhal number St = $(fD)/V$
Re Reynolds Number $(DV\rho)/\mu = (4Q\rho/\pi D)/\mu$

Units

Np nepers
dB decibels

21.1 Introduction to Medical Acoustics

Medical acoustics covers the use of sub-audio vibration, audible sound and ultrasound for medical diagnostics and treatment. Sound is a popular and powerful tool because of its low cost, noninvasive nature, and wide range of potential applications from passive listening to the application of high-energy pulses used for the destruction of kidney stones.

There are four frequency ranges of sound and three levels of intensity that will be presented in this chapter

Table 21.1 Medical frequencies

Title	Period	Frequency	Power/Intensity
Generation	25 years		
Ovulation	28 days		
Diurnal	1 day		
Respiration	6 s	0.15 Hz	
Heart Rate	1 s	1 Hz	
ECG surface bandwidth	10 ms	100 Hz	
ECG intramyocardial BW	250 µs	4 KHz	
Tremor	100 ms	3–18 Hz	
Electromyography surface	5 ms	200 Hz	
Electromyography intramuscle	100 µs	10 kHz	
Flicker fusion (visual)	60 ms	16 Hz	
Vibration sensation	10 ms	100 Hz	
Sound low	50 ms	20 Hz	
Telephone sound speech	3–0.3 ms	0.3–3 kHz	
Sound high	50 µs	20 kHz	
Electro-encephalography (EEG) surface waves			
Delta waves	300 ms	< 3.5 Hz	
Theta waves	200 ms	3.5–7.5 Hz	
Alpha waves	100 ms	7.5–13 Hz	
Beta waves	50 ms	13–30 Hz	
Intracerebral EEG	100 µs	10 kHz	
Ultrasound cleaners		42 kHz	100 W, 1 W/cm^2
Ultrasound bacteriocidal		22.5 kHz	
Ultrasound therapeutic	0.5 µs	1 3 MHz	15 W, 3 W/cm^2
Ultrasound diagnostic			
Transcranial		1–3 MHz	100 mW/cm^2
Heart, abdomen		2–5 MHz	100 mW/cm^2
Vascular, skeletal		4–10 MHz	100 mW/cm^2
Skin, eye		10–40 MHz	8 mW/cm^2
Ultrasound high intensity focused (HIFU)		2–6 MHz	2000 W/cm^2
Diathermy, shortwave, physiotherapy		13–40 MHz	2–15 W
Diathermy, shortwave, cell destruction		300 kHz	100 W
Diathermy, cancer ablation		500 kHz	500 W
Diathermy microwave		0.4–2.5 GHz	
Light		10^{14} Hz	
X-ray		1×10^{18} Hz 4×10^{18} Hz	

1. Infrasound, acoustic waves with frequencies below the level of human hearing. These vibrations, called *thrills*, are often caused by turbulent blood flow (which is abnormal). At the lowest extreme, the pulse can be considered infrasound with a frequency below 2 Hz. With fingertips, the physician can feel vibrations with amplitudes of 1 μm.
2. Audible sound, acoustic waves with frequencies in the range of human hearing. The audible frequency range for the average human is approximately 20 Hz–20 kHz, with a peak sensitivity around 3–4 kHz. At 1 kHz, the standard reference frequency, the ear is able to detect pressure variations on the order of one billionth of an atmosphere. This corresponds to intensities of about 0.1 fW/cm^2 (10^{-16} W/cm^2 = 10^{-12} W/m^2). The oscillatory velocity of air molecules carrying such sound intensities is $v = 50$ nm/s (1 nm = 10^{-9} m) and the maximum pressure fluctuation is 20 μN/m^2 or 0.2 nanoatmospheres (nanobar). At 1 kHz, the air molecules oscillate back and forth about 0.05 nm. The intensity of conversation is 0.1 nW/cm^2 (1 μW/m^2) causing molecules to oscillate with amplitudes of 50 nm, and oscillatory pressures of 20 mN/m^2.
3. Low-frequency ultrasound, which is used for cleaning and tissue disruption.
4. Radio-frequency ultrasound, acoustic waves with frequencies near 1 MHz and wavelengths in the body near 1 mm.

A list of interesting frequencies in medicine is provided for reference Table 21.1.

21.2 Medical Diagnosis; Physical Examination

The most valuable application of acoustics in medicine is the communication from a patient to their physician. In the modern world of *efficient medicine*, however, physicians are encouraged to avoid wasting time listening to the patient so that objective tests can be ordered without delay. This approach avoids physician frustration from sorting out the complaints and symptoms of the patient, even though the patient initiated the appointment with the physician.

Having gathered background information, the physician can proceed to the physical examination, which includes optical methods, olfaction, palpation, testing the mechanical response of tissues and systems to manipulation, percussion, remote listening and aided auscultation, and acoustic transmission and reflection techniques. Only those techniques related to medical acoustics will be discussed further here. Of future importance in medical practice though, are those methods that can be performed remotely, via the internet.

21.2.1 Auscultation – Listening for Sounds

Sounds in the body are generated through the presence of disturbed flow, such as eddies and turbulence, by frictional rubs, and/or by other sources with high accelerations. There are both normal and abnormal sounds within the body. Auscultation is the act of listening for sounds within the body for diagnostic purposes. This may be conducted on any part of the body from head to toe (Fig. 21.1). Anesthesiologists even use esophageal stethoscopes for convenient acquisition of chest sounds. Information gathered during auscultation includes the frequency, intensity, duration, timing, and quality of the sound.

Lung. Both normal and abnormal sounds can be present in the lungs. The source of normal lung sounds is turbulent flow associated with air movement through orifices along the airways. These are large, low-frequency oscillations that cause the surrounding tissues to vibrate within the audible frequency range. The respiratory system is naturally optimized to minimize the work needed to exchange air, so little energy is converted into acoustic power, resulting in low-amplitude sounds. With constriction of the airways, in such conditions as asthma, the air velocity through the airways increases (conservation of mass). Through conservation of energy and momentum, an increase in flow velocity results in an increase in pressure loss, raising the power needed to drive the system. The energy is of course not lost, but converted, in part, into acoustic power. The increase in flow velocity results in an increase in *turbulent diffusion*, or the transfer of energy from the flow to the surrounding tissues through a growing chain of turbulent eddies. When the passages are smaller, the eddies decrease in size, causing an increase in the frequency of the audible sound. The pressure drop caused by a small lumen is higher, causing an increase in eddy velocity and therefore an increase in acoustic power, and the flow rate is decreased,

Amazon EX1014 Page 847

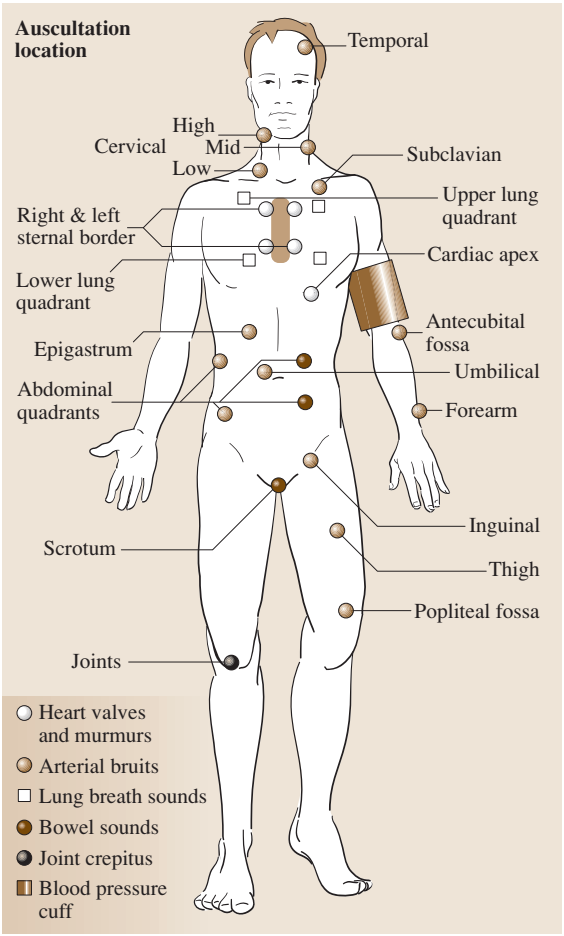


Fig. 21.1 Auscultation sites. Arterial bruits can be heard bilaterally indicating arterial stenoses. Heart valve sounds and heart murmurs are primarily heard in the peristernal intercostal spaces indicating cardiac timing, and valve stenosis and incompetence (insufficiency or regurgitation). Breath sounds in the lung indicate both airway sufficiency, lung inflation, alveolar inflation and pleural friction. Bowel sounds [21.1,2] indicate normal peristalsis; bowel in the scrotum indicate intestinal hernia. Crepitus on motion indicates air in the tissue, bone fracture or joint [21.3–6] cartilage damage. Auscultation in the antecubital fossa during blood pressure cuff deflation yields Korotkoff sounds [21.7,8], 1 = a sharp sound each systolic peak when the cuff pressure equals the systolic pressure indicating the momentary separation of the artery walls and the onset of blood flow, 2 = a loud blowing sound due to blood turbulence, 3 = a soft thud when the artery under the cuff is exhibiting maximum change in arterial diameter, 4 = a soft sound indicating slight arterial deformation near diastolic pressure, 5 = silence when the cuff pressure is below the diastolic arterial pressure

causing a longer duration of flow to fill and empty the lungs. Thus the breath sounds are louder, longer, higher-pitched sounds compared with normal lung sounds. In some medical conditions, the flow dynamics may be just right to produce coherent eddy structures, such as one would see in a smoke stack. These signals result in a narrow-band frequency tone.

Abnormal snapping or popping sounds may also be present in the lung. In the normal lung, the alveoli (terminal air sacks, 280 μm in diameter, 3×10^8 in lung) are filled with air. In pneumonia, regions of alveoli are filled with fluid, which prevents a rapid increase in size without the introduction of a gas phase. The rapid introduction of a gas phase when the lung is expanded during the inspiration phase of respiration causes a tiny snapping sound (broadband acoustic impulse) in each alveolus; many snaps together cause the sound of *rales*.

Table 21.2 Diagnosing valve disease from heart murmurs

	Time	Side	Location	Position
Tricuspid stenosis	Diastolic	Parasternal	3rd ICS	Supine
Tricuspid regurgitation	Systolic	Peristernal	3rd ICS	Supine
Pulmonic stenosis	Systolic	Right	Superior	Supine
Pulmonic regurgitation	Diastolic	Right	Superior	Seated
Mitral stenosis	Diastolic	Left	Apex	Supine
Mitral regurgitation	Systolic	Left	Apex-Axilla	Supine
Aortic stenosis	Systolic	Parasternal	Superior	Supine
Aortic regurgitation	Diastolic	Parasternal	Superior	Seated
Side parasternal = in the intercostal spaces on either side of the sternum				
Location 3rd ICS = the space between the 3rd and 4th rib, Erb's point				
Position seated = Seated, leaning forward, breath in expiration				

Amazon EX1014 Page 848

Because of the lubricant normally present in the pleura located between the lungs and the chest wall, there are normally no sounds generated through a friction rub due to the sliding of the lung along the chest wall. When the pleura are inflamed, coarse rubbing sounds can be heard during breathing.

Cardiac. The cardiovascular system also generates normal and abnormal sounds that can be heard in the chest with a stethoscope [21.9] (Table 21.2). Normal heart sounds [21.10–12] are caused by rapid blood decelerations resulting from the closure of the heart valves. Each of the four heart valves, the tricuspid, pulmonic, mitral and aortic (in order from the venous circulation to the arterial circulation), are flaccid sheets of tissue securely tethered in the closed position to form check (one-way) valves. Like a parachute opening, the valve leaflets move with the reversing blood flow until the tethers become taught, then the leaflets force a rapid deceleration of the blood which generates a sudden (broadband) *thump* with an acoustic power of several watts. No similar rapid acceleration is associated with valve opening. Because the sounds are primarily radiated along the axis of acceleration and conduct better through tissue than the air-filled lungs, the sound of each valve is most prominent at a particular location on the chest wall.

Though four valves are present, usually only two thumps are heard, and they occur at vary precise times in the cardiac cycle. 20 ms after the onset of ventricular contraction, the tricuspid and mitral valves close, causing the first heart sound. 50 ms after closure of the atrioventricular valves, the pulmonic and aortic valves open silently. 320 ms after the onset of ventricular contraction, when ventricular ejection is complete and the ventricular pressures drop below the pulmonary trunk and aortic pressures, the pulmonic and aortic valves snap shut, generating the second heart sound. These times do not vary with heart rate; each cardiac cycle acts like an independent impulse-response system. The systolic ejection volume does vary with the diastolic filling period from the prior cardiac cycle.

Thus, each of the two heart sounds is generated by two valve closures, one in the left heart and one in the right heart. These usually occur simultaneously, however, the sounds do split into two if the inflow valves (tricuspid and mitral) or outflow valves (pulmonic and aortic) close at different times. These *split heart sounds* [21.13] can be indicative of an abnormality. The closure of the tricuspid valve is dependent on contraction of the right ventricle, which in turn is dependent on the electrical conduction system from the pacemaker of the heart. As a result, a delay in conduction or a slow response of the right ventricular muscle (myocardium) will delay the tricuspid sound, causing splitting. Similarly, if there is a delay in the conduction system of the left heart, or the left ventricular myocardium contraction is slow, then the mitral sound is delayed. The closure of the outflow valves is dependent on contraction strength, pressure in the outflow vessels, and ejection volume. A mismatch of any of those factors will cause the pulmonic and aortic valves to close at different times, splitting the second sound. Even the effects of respiration on the filling and emptying of the heart chambers cause some normal heart sound splitting.

Abnormal sounds that occur in the cardiovascular system are called *bruits* and *murmurs* [21.14–17]. A murmur is a sound lasting longer than 10 ms heard by a cardiologist from the heart; a bruit is a similar sound, but heard by an angiologist from an artery. Veins rarely make sounds; venous sounds are called *hums* because they are continuous. A *thrill* is a similar vibration felt by the hand. Because these represent power dissipation from the cardiovascular system, all are considered abnormal. The absence of normal bruits and murmurs results from a biological feedback that forms the shape of the cardiovascular system. The feedback mechanism is not well understood. In fluid mechanics, the transition from laminar flow to turbulent flow occurs near a Reynolds number of 2000. Turbulent flow is highly dissipative while laminar flow is minimally dissipative. Coincidentally, peak flow through the cardiovascular system maintains a condition of Reynolds number less than 2000. If a heart valve or an artery is required to carry higher flow rates, then the conduit will dilate until the Reynolds number (proportional to the flow divided by the diameter) becomes less than 2000. Distal to a stenosis, where turbulence occurs, the conduit dilates, causing post-stenotic dilation. Thus, it appears that sustained turbulence, which dissipates power, is the driving force for remodeling to avoid turbulence and make the hemodynamic system more energy efficient.

The vibrations indicate the presence of disturbed blood flow, caused by deceleration of the blood beyond a narrow stenosis. The vibrations are associated with a pressure drop across the stenosis because of a transfer of hydraulic power to acoustic power through a cascade of eddies. In turbulent flow, the vibrations are broadband. When coherent eddies are present, the vibrations are narrow band, and the frequency of the eddies can be related to the flow velocity (v_f) and vessel diameter (d) through the Strouhal number ($\mathrm{St} = fd/v_f$). In addition to the location and power of a bruit, the frequency of the bruit is also helpful in evaluating its clinical im-

portance. The Strouhal number has been used for the prediction of the residual lumen diameters of carotid stenoses. If the frequency of a bruit is divided into the number 500 [mm Hz], the result is the minimum lumen diameter [21.18–25] in mm. In clinical testing, this method accurately predicted the stenosis in hundreds of cases. However, because cumbersome technology is involved, and because the method fails in tight stenoses, which are most clinically important, this clever method has been abandoned.

Heart murmurs are auscultated at a series of locations on the precordium (the front middle of the chest around the sternum). They are graded from grade 1 (barely audible) to grade 6 (no stethoscope needed). Murmurs may result from leaky (incompetent, regurgitant) valves, stenotic (narrow) valves, or other holes between heart chambers. The differential diagnosis of heart murmurs is based on the timing of the murmur and the locations on the chest wall [21.26, 27] where the murmur is best heard. A murmur can be early systolic, mid-systolic, late systolic, or early diastolic, or mid-diastolic. If a murmur is best heard in the left second intercostal space (the space between the second rib and the third rib along the left boarder of the sternum), it is most probably an aortic stenosis [21.28–30]. Each of the other valve pathologies has a corresponding location where the murmur is likely to be highest. Of course, there is a great deal of variability between patients, and the examiner must apply a full set of detective skills to sort out a proper diagnosis.

There are some common systolic heart murmurs. Athletes, with temporary increased blood flow due to exercise programs have physiologic systolic heart murmurs, as do pregnant women, who have a 30% increase in cardiac output to supply the placenta. Hemodialysis patients also often have systolic heart murmurs because of the increased cardiac output required to supply the dialysis access shunt [21.31] in addition to the supply to the body. In these cases, remodeling of the blood vessels to increase the flow diameter has not had time to occur.

Bruits are almost always associated with stenoses in arteries, which often require treatment. They are classified by their pitch and duration. The occurrence of bruits in the head, in the abdomen after eating and in the legs are all indications of arterial stenoses, or high-velocity blood entering vascular dilations. The presence of a bruit heard in the neck is usually indicative of a carotid stenosis. The carotid arteries are the major vessels located on the left and right side of the neck that supply blood to the brain. The development of an atherosclerotic plaque causes luminal narrowing and the corresponding onset of disturbed flow. This lesion is associated with a 20% risk of stroke in two years from material that can be released by the atherosclerotic plaque and travels to the brain to occlude a branch artery.

Carotid bruits are usually present when the residual lumen (stenotic) diameter is greater than 50% of the original lumen diameter (called a 50% diameter reduction), but less than 90% of the original lumen diameter (90% diameter reduction) In this range, the blood flow rate to the brain through the carotid artery is approximately 5 cm^3/s. If the pressure [21.32] drop across the stenotic region is about 20 mmHg, then the power dissipation at the stenosis is 13 mW, which can appear as sound energy in a bruit. Unfortunately, as the stenosis becomes more severe, the flow rate along the artery decreases, reducing the acoustic power for the bruit, so the bruit is no longer detectable, even though this more severe stenosis carries a higher risk for stroke. The decreased blood flow will not cause symptoms in most people because 95% of the population have collateral connections (the circle of Willis) that can provide the required blood flow that is not delivered by the stenotic artery. In the remaining 5% of the population, a severe carotid stenosis causes impaired mental function.

One of the barriers to new methods in the classification of carotid stenosis is that the first description of the relationship between carotid artery stenosis and apoplexy (stroke) was provided by Egaz Moniz in 1938. Moniz used percent diameter reduction and everyone has followed suit, through a confusing and contentious evolution, even when more-rational and useful alternatives have been developed. Thus, a new method that predicts residual lumen diameter will not be easily adopted by the medical community because of the tradition of using percent diameter reduction.

Bruits may be generated in the abdomen by stenoses in the renal arteries, which supply the kidneys. A renal artery stenosis will cause renovascular hypertension because the kidney is a transducer for detecting low blood pressure. The pressure drop across a renal artery stenosis causes the kidney to measure a low pressure, and deliver the signals to raise blood pressure in the arterial system, causing hypertension in all arteries except those in the affected kidney. The transduction system is called the renin-angiotensin-aldosterone system.

Some cardiovascular sounds are so loud that they cannot be missed. A patient in renal failure, refusing dialysis, had a cardiac friction rub (friction in the pericardial sack from inflammation). The 1 kHz *squeak* occurring with each systole could be heard from the doorway of the patient's room.

A stethoscope can also be used to detect bowel sounds from the abdomen. These sounds are generated by the liquid and gaseous contents of the bowel being forced from one intestinal segment to another through an orifice created by peristaltic action. The same fluid eddies described above are responsible for these sounds. The presence of bowel sounds is an important indicator of a normally functioning gut. It is common, the day after abdominal surgery, to check to be sure that normal bowel sounds are present.

Palpation, though not an acoustic technique, is used to feel, rather than hear, the same vibrations. Hemodialysis patients have a *vascular access* shunt created by tying an artery directly to a vein. This shunt may have a flow rate of 3 liters per minute, half of the cardiac output. The pressure drop is equal to the systolic blood pressure 120 mmHg. So the hemodynamic power dissipated in the shunt is 800 mW of power. This energy is converted to a vibration (thrill) which can be felt with the fingers.

Acoustic Stethoscope

Auscultation may be performed directly, but is most commonly conducted with a stethoscope. The device was invented by Dr. Rene Theophile Hyacinthe Laennec in 1816. While examining a patient, Dr. Laennec rolled up a sheet of paper and placed one end over the patient's heart and the other end over his ear. Dr. Laennec later replaced the rolled paper with a wooden tube (similar in appearance to a candlestick), which was called a stethoscope from the Greek words *stethos* (chest) and *skopein* (to look at). Two modern versions of the stethoscope are notable: the binaural stethoscope and the fetal stethoscope. The common binaural stethoscope consists of a bell connected by flexible tubing to a pair of earpieces. Each earpiece is sealed into the corresponding external auditory meatus. The bell is sometimes covered with a diaphragm, which acts as a high-pass filter. For example, normal heart sounds have frequencies near 50 Hz while heart murmurs have frequencies exceeding 100 Hz. Filtering out the normal heart sounds with the diaphragm of the stethoscope often allows pathologic murmurs to be detected. When using an open bell stethoscope, a diaphragm can be formed by pressing the bell tighter to the skin, stretching the skin and deeper tissue, which attenuates the lower frequencies [21.33–36].

The fetal stethoscope is designed for use in the auscultation of fetal heart sounds, which can be detected during the final half of a 40 week pregnancy. The fetal stethoscope uses a rod connecting the mother's abdominal wall to the examiner's forehead to transmit the sound by bone conduction in the examiner's head to the inner ear where it is converted into nerve impulses. This avoids the impedance changes between tissue and air at the skin surface and then again from air to tissue in the middle ear.

For the last 20 years, small ultrasonic Doppler ultrasound stethoscope systems have been available to detect the fetal heart beat [21.37, 38]. They operate like the ultrasound imaging *phono-angiograph* to easily detect solid tissue vibrations from deep in the body. These continuous-wave instruments allow the pregnant mother to perform easy self-examination. They are available on a rental basis without a medical prescription.

Electronic Stethoscope

Electronic stethoscopes [21.39–41] provide improved amplification and background noise cancelation over traditional stethoscopes. Some also have recording capabilities to provide quantitative, rather than just qualitative, assessment of the sounds heard. For example, by listening to the neck with a phono-angiograph (Fig. 21.2), the location and duration of the bruit can be documented. A phono-angiograph is just an electronic stethoscope with an attached oscilloscope. If applied to the heart, then it is called a phono-cardiograph. This bruit (Fig. 21.2), is a pan-systolic bruit (lasting all of systole) but not a pan-diastolic bruit (lasting all of systole and diastole). Since arterial blood velocities are higher

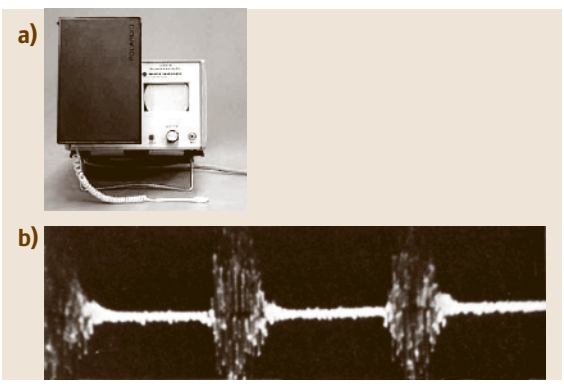


Fig. 21.2a,b Carotid phono-angiography (CPA) **(a)** A microphone fitted with an acoustic coupling to the neck gathers sounds, which are displayed on an oscilloscope and captured with an instant camera. **(b)** The sounds displayed are classified by the location, amplitude, and duration of the bruit. A similar method, phono-cardiography, was used for the heart. A method, quantitative phono-angiography (QCPA) provided an analysis of the frequency content of the bruit

Amazon EX1014 Page 851

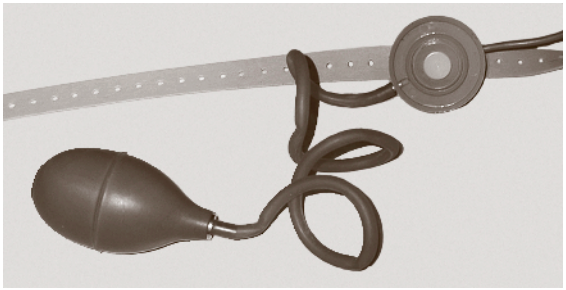


Fig. 21.3 Phono-cardiograph transducer. To avoid sounds and motions introduced by the examiner's hand on the chest, this phonocardiography transducer is fitted with a suction ring and elastic strap to secure the transducer element to the chest. Early echocardiographs had provision for amplifying and displaying the phono-cardiograph signals, which were filtered to display audible frequencies or infrasonic frequencies

during systole than diastole, all arterial bruits include systole. The exceptions, where arterial flow in a collateral pathway is reversed due to a *steal syndrome* are so rare that common names have not been adopted for this condition. To be sure that the bruit is systolic, the examiner can palpate the pulse at the wrist while listening, to associate the bruit timing with systolic arterial motion.

However, if such an image were gathered with a phono-cardiograph (Fig. 21.3) from the heart, it would be impossible to know whether the murmur was coming from aortic or pulmonic stenosis (during systole), aortic or pulmonic regurgitation (during diastole), from tricuspid or mitral regurgitation (during systole), or tricuspid or mitral stenosis (during diastole), unless there were some indication of the timing of the cardiac cycle. A phono-cardiogram displays an associated electro-cardiograph (ECG) tracing to provide that timing information.

Unfortunately, electronic stethoscopes, phono-angiographs and phono-cardiographs are considered too complicated for use in everyday medical practice. Although they do provide a graphical output for the patient chart (required documentation to receive payment for diagnostic procedures), a separate current procedural terminology (CPT) code is not available for phono-angiography, phonocardiography or for stethoscope examination. Like the stethoscope examination, these procedures, if performed, are considered part of a standard office visit. Physicians are reluctant to use technical examination methods if they cannot charge for the extra time involved. Although a few research facilities documented bruits and murmurs detected with these systems between 1965 and 1985, the systems have nearly disappeared from medical practice. Like the first half of the 20th century, physicians today are trained to use acoustic stethoscopes. Of course, as physicians become more experienced and more mature, their hearing deteriorates. Even the most respected cardiologist will fail to correctly diagnose cardiac valve disease in 50% of patients. Bruits with 35 μm, 300 Hz vibrations from the jugular vein (Fig. 21.4) documented by a new ultrasound imaging phono-angiograph were not detected by stethoscopic auscultation. So, in spite of the ubiquitous use of stethoscopes for bruit/murmur detection, their sensitivity may not be adequate for the task.

21.2.2 Phonation and Auscultation

In addition to listening for sounds with a stethoscope, an examiner may cause the generation of sounds to assess transmission or resonance. During auscultation of the lungs, an examiner may ask the patient to phonate (make voice sounds) at different frequencies "Say 'eeeeeee', now say 'aaaaaaaa'". During that period, the examiner will either palpate the chest wall for vibration or listen to the chest wall with a stethoscope. Poor sound transmission may indicate that the airways are obstructed, enhanced sound transmission (enhanced *fremitus*) indicates fluid has filled a space normally occupied by air.

21.2.3 Percussion

Percussion is another method of sound generation by striking the chest or abdomen to assess transmission or resonance. The skin may be struck with the hand or finger, or one hand may be placed on the skin and struck with the other hand. A higher-frequency shorter pulse can be generated by placing a coin on the skin and striking the coin with another coin. The resultant sound can be assessed by listening to the resonance through the air or by using a hand or stethoscope to assess transmission of the impulse. Percussion that results in a dull sound indicates a fluid-filled space below the skin; a resonant sound indicates an air-filled space below the skin. These methods are often used to find the lower boarder of the lungs, to identify fluid-filled consolidation in the lungs, and to find the edge of the liver (fluid-filled) in the abdomen (air-filled intestines).

These acoustic examination techniques have been developed in medical practice over centuries, and although efforts have been made to use electronic aids for

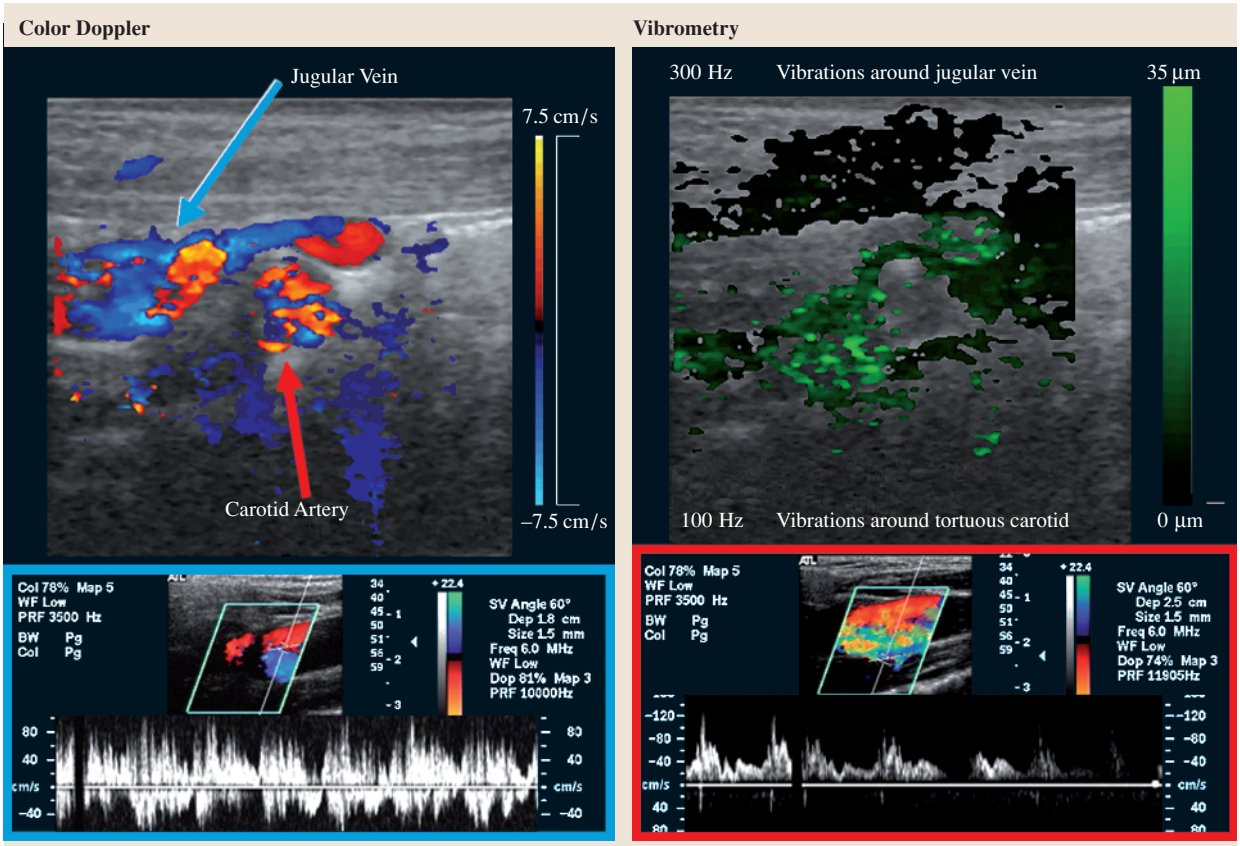


Fig. 21.4 Imaging vibrometry of the neck. Color Doppler image (*upper left*), vibration image (*upper right*), jugular vein Doppler (*lower left*) and carotid artery Doppler (*lower right*) of a patient complaining of "vibrations in the head". The 300 Hz vibrations with an amplitude of 35 µm could not be heard with a stethoscope by either of two experienced examiners (Courtesy M. Paun, S. Sikdar)

diagnosis and for training, still, the common practice is to learn these methods from a mentor. In the last quarter of the 20th century, ultrasound has become adopted in specialty medical practice, but has not made inroads in general practice. It is likely that, if and when ultrasound diagnostic methods become widely used in general medical practice, it will be used with the same level of skill as stethoscope auscultation.

21.3 Basic Physics of Ultrasound Propagation in Tissue

The vibrations of molecules at any location (or time) in a sound wave can be characterized by four parameters: the pressure fluctuation from the mean static pressure (p), the molecular velocity (v), the displacement of the molecules from the resting location (d), and the molecular acceleration (a). The two most basic parameters, pressure deviation and molecular velocity are dependent only on the intensity of the sound. Their ratio, p/v, is determined by the acoustic impedance ($Z = p/v$), a property of the tissue. At the sound level shown in Fig. 21.5, 300 mW/cm^2 (3 kW/m^2), the maximum acoustic pressure fluctuation equals 1 atm. If the sound intensity is increased, then the negative peaks represent a pressure below zero. Most ultrasound examinations are performed at pulse peak intensities 25 times this high, and therefore, peak amplitudes five times atmo-

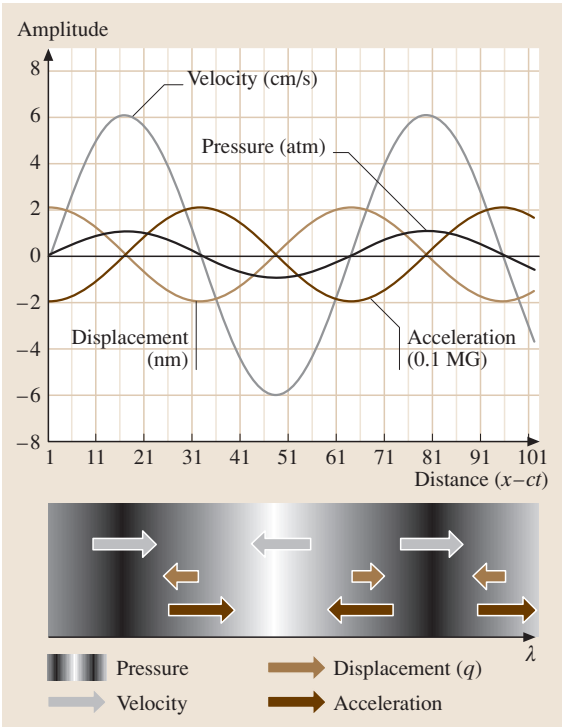


Fig. 21.5 Typical acoustic parameters of a 300 mW/cm^2, 5 MHz ultrasonic wave in tissue impedance. Input parameters: Tissue impedance = $Z = 1.5$ MRayles = 1.5 MN s/m^3 = 1.5×10^6 kg/(m^{2}s), pressure amplitude = $p = 1$ atm (bar) = 101 kPa (kN/m^2) = 760 mmHg (Torr), frequency = 5 MHz; resultant wave parameters: intensity = 0.33 W/cm^2 = $p^2/2Z$, peak molecular velocity = 6 cm/s = p/Z, molecular displacement = 1.9 nm, molecular acceleration of 0.19 million times the acceleration of gravity

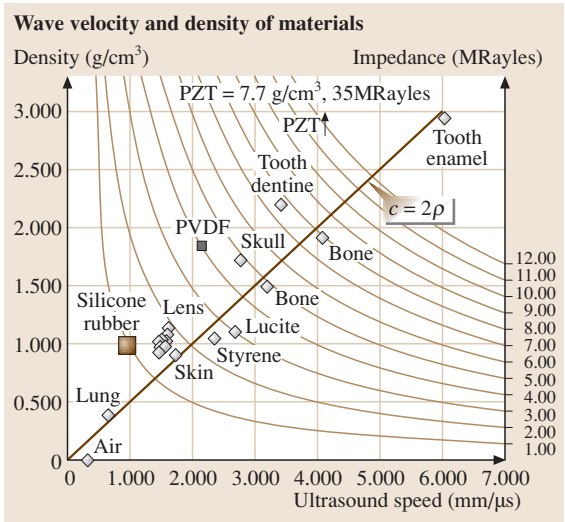


Fig. 21.6 Acoustic properties of materials of medical interest. Empirically, the acoustic properties of tissues fall on a line relating ultrasound speed (c) and tissue density (ρ); $c = 2\rho$. Using the equation for impedance $Z = \rho c$, then $Z = c^2/2 = 2\rho^2$. Polyvinylidene fluoride (PVDF) transducer material has acoustic properties near soft tissue, lead zirconate titanate (PZT) transducer material has much higher density and acoustic impedance

spheric pressure, making the compressions 6 atm and the decompressions −4 atm.

The linear acoustic relationships can be easily derived using a simple linear elastic model of tissue (Fig. 21.5) and Newton's first law of motion on the molecules. The molecules in the tissue will accelerate along the axis of the ultrasound propagation (x) and displace a distance (q) away from equilibrium because of a gradient in pressure ($\mathrm{d}p/\mathrm{d}x$), which is related to the tissue compression ($\mathrm{d}q/\mathrm{d}x$) and the tissue stiffness (κ).

$$p = -\frac{\kappa\,\mathrm{d}q}{\mathrm{d}x} = -\kappa q' \,, \tag{21.1}$$

$$\kappa\frac{\mathrm{d}^2 q}{\mathrm{d}x^2} = -\frac{\mathrm{d}p}{\mathrm{d}x} = \rho\frac{\mathrm{d}q^2}{\mathrm{d}t^2} \,. \tag{21.2}$$

Solving this equation with the variable $q(x - Ct)$ produces

$$\kappa q'' = \rho C^2 q'' \,, \tag{21.3}$$

so

$$C = \sqrt{\frac{\kappa}{\rho}} \,. \tag{21.4}$$

The molecular velocity ($v = \mathrm{d}q/\mathrm{d}t = -Cq'$) and the tissue pressure ($p = -\kappa q'$) form an important ratio

$$Z = \frac{p}{v} = \frac{\kappa}{c} = \sqrt{\kappa\rho} = \rho C \,. \tag{21.5}$$

Although the tissue density (ρ) and stiffness (κ) determine the acoustic impedance (Z) and acoustic wave speed (C), it is more convenient to measure the density and wave speed of a material and compute the impedance.

A good deal of effort has been devoted to measuring the properties of tissues (Figs. 21.6–21.8). Measurement differences between laboratories and differences in tissue preparation contribute to the variability of tissue property values. While living mammalian body tissues

Amazon EX1014 Page 854

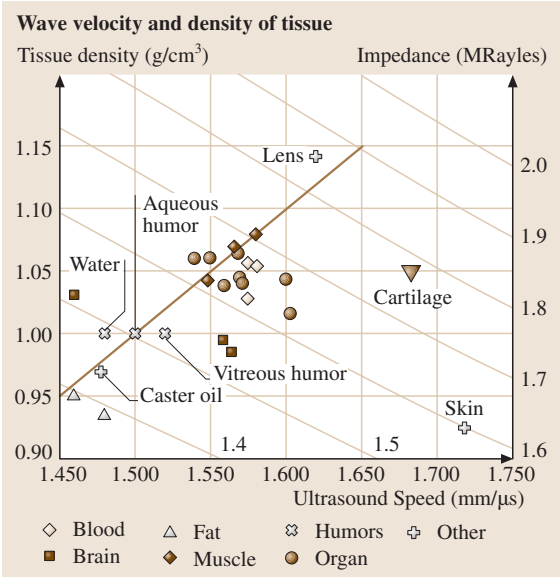


Fig. 21.7 Acoustic properties of soft tissues

are held to a temperature near 37 °C, and have blood flow that swells the tissue and changes the average composition with the heart rate and respiratory rate, it is convenient to measure properties of dead tissue at bench-top temperatures (25 °C). Fortunately, the exact acoustic properties of tissues have not proven very useful in clinical diagnostic applications. Ultrasound imaging methods have not been able to differentiate tissue types in spite of a half century of efforts. The identification of disease with ultrasound imaging uses the shapes of large structures as the primary diagnostic tool.

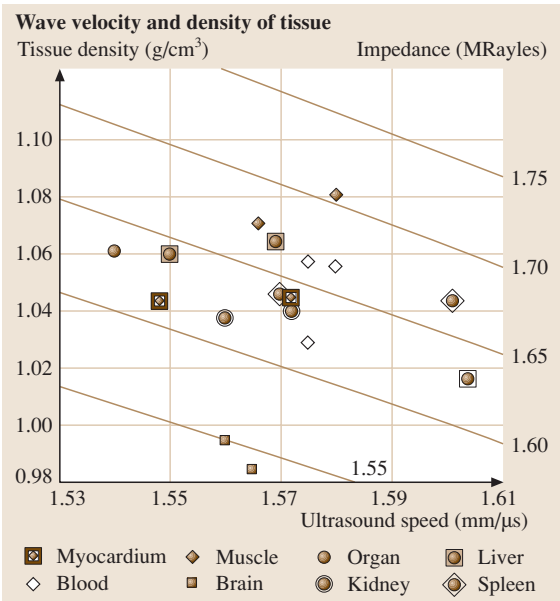


Fig. 21.8 Detailed acoustic properties of soft tissues

21.3.1 Reflection of Normal-Angle-Incident Ultrasound

When ultrasound intersects an interface between two materials with differing acoustic impedances (Z_1 and Z_2), perpendicular to the interface, then the amplitude of the reflection of ultrasound at the interface can be computed from the requirement that the pressure change and molecular velocity of the incident wave (p_i, v_i) must equal sum of the pressure changes and molecular velocities of the transmitted and reflected waves (p_t, p_r, v_t, v_r). The impedances of the materials relate the pressures and molecular velocities ($Z_1 = p_i/v_i = -p_r/v_r$, $Z_2 = p_t/v_t$). The pressure reflection and transmission coefficients are

$$\frac{p_r}{p_i} = R = \frac{Z_2 - Z_1}{Z_2 + Z_1} , \tag{21.6}$$

$$\frac{p_t}{p_i} = T = 2\frac{Z_2}{Z_2 + Z_1} . \tag{21.7}$$

Note that if the acoustic impedance of the second material is lower than the acoustic impedance of the first material, then the reflected wave is inverted. The sum of the transmitted and reflected power equals the incident power.

21.3.2 Acute-Angle Reflection of Ultrasound

If the ultrasound is incident on the interface at an acute angle, the ultrasound that passes into the second tissue is tilted to a new direction according to Snell's law (Fig. 21.9).

$$\frac{C_1}{\sin\theta_1} = \frac{C_2}{\sin\theta_2} . \tag{21.8}$$

Because the component of molecular velocity normal to the surface must be preserved:

$$\frac{p_r}{p_i} = R = \frac{\frac{Z_2}{\cos\theta_2} - \frac{Z_1}{\cos\theta_1}}{\frac{Z_2}{\cos\theta_2} + \frac{Z_1}{\cos\theta_1}} . \tag{21.9}$$

In this case, not all of the ultrasound power is found in the sum of the reflected wave and the transmitted wave. The remaining portion of the power appears as a surface *capillary* wave propagating along the surface of

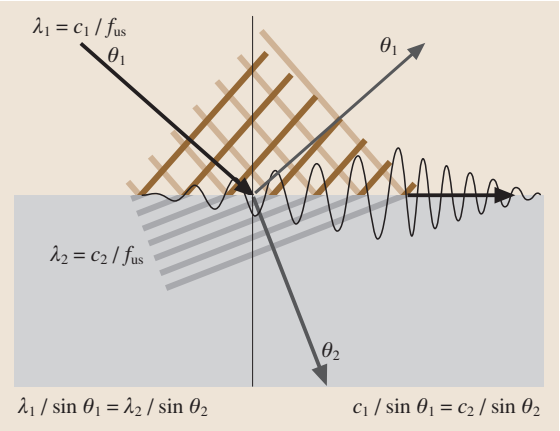


Fig. 21.9 Refraction of ultrasound as it passes from one material to another. *Dark brown*: incident compressional wavefronts and wave vector incident at angle θ_1. *Light brown*: reflected compressional wavefronts and wave vector reflecting at angle θ_1. *Gray*: transmitted compressional wave. The interfacial shear wave increases in intensity as the ultrasound beam cross section coherently contributes to the wave; attenuation of the shear wave causes a rapid decrease in amplitude outside the beam

the interface. This capillary wave is subject to high attenuation. Because tissue has an interface on every cell boundary and every tissue boundary, this mechanism is responsible for some of the conversion of incident ultrasound power into heat as ultrasound propagates into tissue. It is important to notice that the propagation speed of the capillary wave is not the same as that of the component of the compressional wave directed along the interface, so the capillary wave is out of phase with the ultrasound wave, contributing to the transfer of energy into the capillary wave. Outside the ultrasound beam, the capillary wave takes on its conventional wavelength (c_s/f), where f is the ultrasound frequency. The propagation speed of the capillary wave (c_s) is proportional to the interfacial tension.

21.3.3 Diagnostic Ultrasound Propagation in Tissue

Diagnostic ultrasound uses ultrasound frequencies of 1–30 MHz. (MHz = cycles per μs). The speed of ultrasound in most tissues is near 1.5 mm/μs (1500 m/s). There are some notable exceptions. In bone the speed is between 3.5 mm/μs and 4 mm/μs, and in air the speed is 0.3 mm/μs. In cartilage the speed is 1.75 mm/μs which of course varies with composition. In fat the speed is 1.45 mm/μs (Table 21.3). In muscle, the speed of ultrasound varies by about 1.5%, depending on the angle of propagation to fiber orientation.

Although most authors focus attention on ultrasound frequency, ultrasound wavelength may be more convenient for understanding the important issues. There are three issues of importance:

- Ultrasound cannot resolve structures smaller than the wavelength.
- Scattering and attenuation [21.42] both depend on ultrasound wavelength and its relationship to the size of the scatterers.
- The focusing properties of ultrasound beam patterns and the formation of side-lobes depend on the ratio between the wavelength and the width of the transducer.

21.3.4 Amplitude of Ultrasound Echoes

Red blood cells (erythrocytes) are 0.007 mm in diameter and about 0.001 mm thick. The largest cells in the body, nerve axons, may be 1 m long, but their diameter is about 0.001 mm. A liver cell is about 0.04 mm in diameter. The smallest structures imaged by ultrasound in clinical practice are layers less than 1 mm thick, such as the media lining major arteries (the intima-media thickness, IMT) [21.43], corneal layers (with 30 MHz ultrasound) and skin surfaces. However, these structures can only be observed in the range (depth) direction because resolution is best in that direction. Lateral resolution is dependent on the numeric aperture of the ultrasound beam pattern and the relationship between the object of interest and the focus as well as the ultrasound wavelength. Only the ultrasound wavelength affects depth resolution. The result is that ultrasound images are anisotropic, with much better resolution in the depth direction than in the lateral direction. The image is delivered to the examiner as a pictorial map in two dimensions. In our optical experience, resolution is equal in the two dimensions. Therefore, we expect that, if we can see detail in one dimension, we can also see the detail in the other, even when we cannot. This gives the impression that ultrasound has higher resolving power in the lateral direction than it in fact does. This is particularly important in measurements such as the IMT and measuring the thickness of the atherosclerotic cap. Such structures can only be visualized when the ultrasound beam is oriented perpendicular to the structure.

To identify tissues and organs, the lateral dimensions must by at least 10 wavelengths wide. This

Amazon EX1014 Page 856

Table 21.3 Ultrasound wavelength versus tissue thickness. To resolve the thickness of these structures, the ultrasound frequency must be much higher than listed in this table. Much lower-frequency ultrasound passes through with little interaction. For particles or cells, lower-frequency ultrasound is subject to Rayleigh scattering

Tissue	Property	Dimension (mm)	Ultrasound	
			Speed (mm/μs)	Frequency (MHz)
Trachea	Lumen diameter	25	0.32	0.013
Aorta	Lumen diameter	30	1.57	0.063
Fat layer	Subcutaneous	20	1.45	0.080
Myocardium	Thickness	10	1.6	0.16
Skull	Thickness	10	4	0.4
Brain	Cortex	3	1.5	0.5
Artery	Wall	0.8	1.6	2
Trachea	Wall	0.6	1.75	2.9
Bone	Cortex	1	4	4
Eye	Cornea	0.6	1.64	2.73
Skin	Dermis	0.3	1.73	5.76
Fat	Cell	0.1	1.45	14.5
Brain	Cell, large	0.1	1.5	15
Skin	Epidermis	0.05	1.73	35
Erythrocyte	Cell	0.007	1.57	225
Bacteria	Cell, large	0.005	1.5	300
Brain	Cell, small	0.004	1.5	375

Notes:
1. Trachea lumen: for phonation below 13 kHz, the trachea diameter is smaller than the wavelength so that the trachea acts like a waveguide, bounded by cartilage with acoustic velocity and impedance higher than the luminal air.
2. Bone cortex and thin skull have thicknesses near the wavelength of transcranial ultrasound and a much higher ultrasound impedance, so the skull can act as an interference filter reflecting some wavelengths and passing others.
3. As the upper limit of most medical ultrasound systems is 15 MHz, small animal research instruments image at frequencies as high as 70 MHz, high enough to resolve the epidermis and large brain cells, but not high enough to image most cells.

number varies depending on the numeric aperture of the imaging system and on the acoustic contrast with the surrounding material. Small round structures such as tendons can be imaged using tricks, such as watching for motion, but most objects successfully imaged by ultrasound are larger. It is important to remember that it is possible to see an object that you cannot resolve. For instance, you are able see stars or lights at night, even though you cannot resolve closely spaced pairs of stars. When the signal from a target is strong, the target can be detected and located, even though a pair of targets cannot be resolved as two. If Doppler demodulation is used, rather than amplitude demodulation (used for B-mode), small vessels with

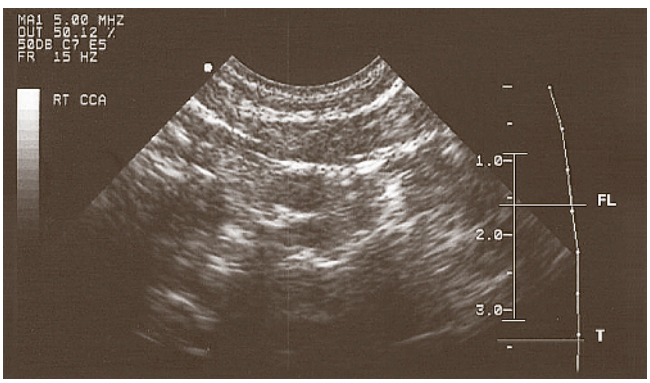


Fig. 21.10 Near-field versus far-field difference in ultrasound image. This is an image taken with a diagnostic scanner using a 5 MHz ultrasound transducer (wavelength = 0.3 mm) with a concavity radius of 16 mm and a transducer diameter of 6.4 mm. The computed *transition zone* is $d^2/4\lambda = 34$ mm. The *focus* must be less than the 34 mm distance to the transition zone to be effective. With a fixed-focus fixed-aperture ultrasound transducer, the transition from the near field to the far field is marked by lateral spreading of the speckle in the far field. There is no effect on the range (depth) speckle dimension

blood flow can be resolved if the signals are strong enough.

Ultrasound imaging is a coherent process that generates speckle in the image. This is often called *tissue texture*, but the spacing of the speckle is about equal to the burst length of the ultrasound pulse, which is up to several wavelengths, and this is much greater than the fine cellular structure of the tissues under observation. Our visual processing is so good that we integrate prior knowledge, and outline shape seamlessly into our recognition process, while believing that we are recognizing texture.

Ultrasound imaging and aeronautical radar imaging are electronically identical. However, in aeronautics, the space between aircraft is very large compared to the size of the aircraft and to the wavelength of the radar wave. In medical ultrasound imaging, the cells are tight packed together, and the cell size (10 μm) is smaller than the wavelength of ultrasound (500 μm) so that the scattering is complex. In addition, ultrasound medical imaging systems often generate images from the complex near field (Fresnel zone) of the ultrasound transducer (Fig. 21.10). This becomes particularly confusing when tissues are moving laterally through an array of ultrasound beams. Because of the Moire effect of this system, tissue motions from left to right can cause speckle patterns to move right to left.

21.3.5 Fresnel Zone (Near Field), Transition Zone, and Fraunhofer Zone (Far Field)

The transition between the Fresnel and Fraunhofer zones can be computed from geometric considerations (Fig. 21.11). As in the computation of the basic wave equations, we assume that ultrasound propagation is a linear process and that superposition holds (pressure fluctuations and molecular velocity fluctuations can be added). Although some authors use Huygens' principle [21.44], dividing the transducer into tiny segments and adding the contribution of each to the wave, the computation is simplified if reciprocity is used (the transmitted ultrasound beam pattern and the received ultrasound beam pattern are the same), and the wave is assumed to be emitted from a test point in the ultrasound beam and received by the transducer.

A wave propagating from a stationary point will expand as a series of concentric spherical surfaces, spaced by the ultrasound wavelength with the source at the center. If the source is close to a planar transducer face, then at every time point some compressional spherical regions and some rarefactional spherical regions will simultaneously be present on the transducer face plane. This provides a mixed signal to the transducer, the compressions and rarefactions destructively inter-

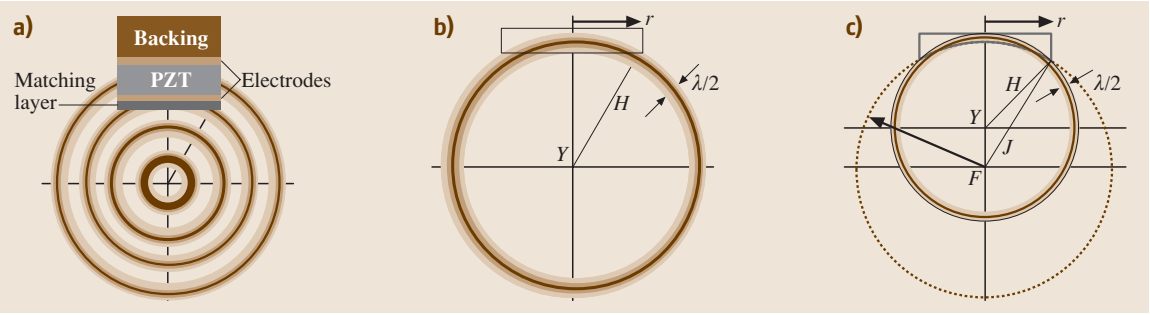
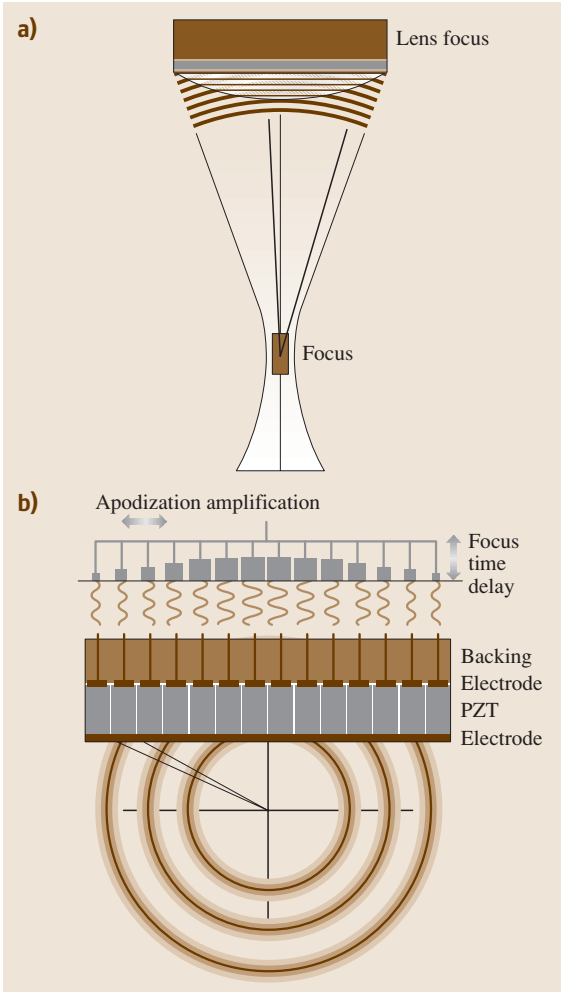


Fig. 21.11a–c Geometry of transition zone between near field (Fresnel zone) and far field (Fraunhofer zone) **(a)** A transducer consists of a piezoelectric element (PZT) with electrodes on the face with an impedance matching layer and on the back with a backing material. For computation, only the wave intersection with the matching layer face needs to be considered. **(b)** By geometry, the distance to the transition zone (Ψ) between the Fresnel zone and the Fraunhofer zone for a flat transducer can be estimated. The transition zone is located at the furthest distance from the transducer that the origin of a compression half-cycle of a wave can completely envelope the transducer face. r = transducer radius, Ψ = distance to transition zone, $\lambda/2$ = half wavelength, H = hypotenuse of transition radius. $H^2 = \Psi^2 + r^2$, $H - \Psi = \lambda/2$. If $\lambda/4 \ll \Psi$, then $\Psi = r^2/\lambda = d^2/4\lambda$, the accepted formula for the transition zone. A wave originating from a closer location (Fresnel zone), will always produce a combination of compressions and decompressions at the transducer surface. **(c)** Drawings can be made for focused transducers whether concave, lens or phased focus array, but the algebra is more complex. The equivalent concave curvature is shown by the *dotted circle*. F = transducer focal radius, J = hypotenuse of focal radius

Amazon EX1014 Page 858

fering. There are some locations of the source point that result in equal areas of compression and decompression of the transducer surface at all times. These are locations where the transducer has no sensitivity, and are called null points. By reciprocity, if the transducer emitted a wave, a null point will have no sound intensity from the transducer. Ultrasound transducer beam plots often show these null locations along the axis of the ultrasound beam. There may be several of these null points along the beam axis, and at other points in the beam pattern. The *near field* or *Fresnel zone* is the region of the ultrasound beam pattern where these null points are present. In the Fresnel zone, there is always some destructive interference of the ultrasound wave. The Fresnel zone is the region where beam focusing can be achieved.



When the source point is at a greater distance from the transducer plane, these regions of compression and rarefaction become nearly planar. There are times when the planar transducer surface can be completely contained in either a compression zone or a decompression zone. At this range, at least near the axis of the ultrasound beam pattern, there are no regions of destructive interference. This is the *far field* or *Fraunhofer zone*.

There is a transition between the Fresnel and Fraunhofer zones along the ultrasound beam axis, where it is first possible for the ultrasound transducer to be just barely completely contained in a compression half-wave region at some time within the ultrasound wave period. Half a period later, the transducer will be completely contained in a decompression half wave. This location is easily diagrammed geometrically (Fig. 21.11b) and the distance to the transition zone computed. Although the geometry for a similar computation for a concave transducer can be easily drawn (Fig. 21.11c), the algebraic solution for the focused transition, which is now retracted toward the transducer, is more difficult.

In addition to using a concave transducer, focusing of the ultrasound beam pattern can also be achieved by the use of a lens on the front of the transducer (Fig. 21.12a). The lens is often made of silicone rubber because of the low ultrasound propagation speed. Currently, most ultrasound scan heads consist of a linear array of 128 transducer elements; each element is 2 cm long and 0.5 mm wide arranged in a side-by-side row with a footprint of 2 cm by 6.4 cm. Focus is pro-

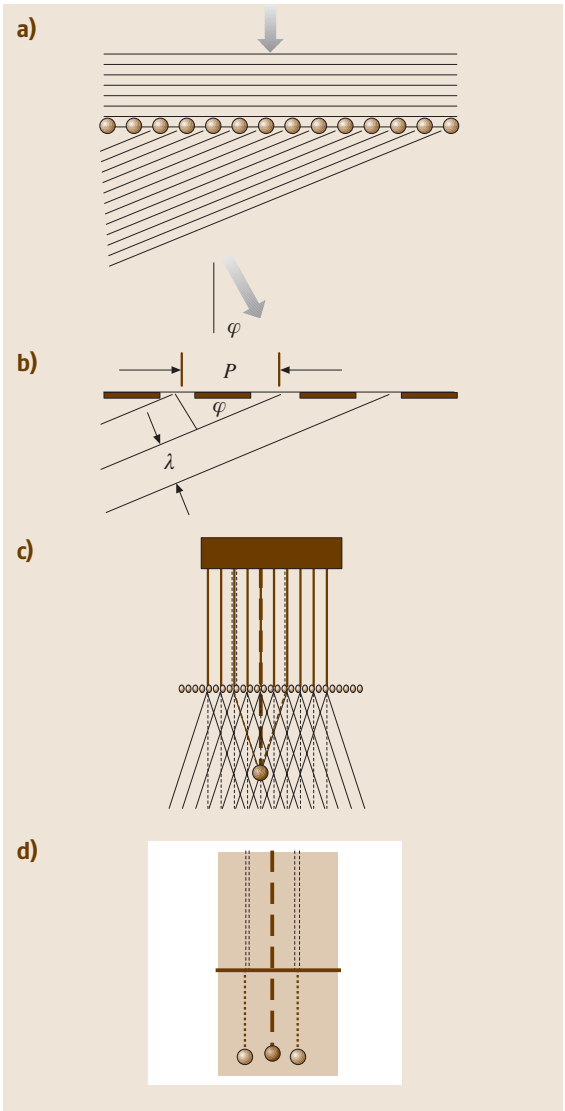
Fig. 21.12a,b Mechanical lens focusing and electronic delay focussing. **(a)** A low acoustic velocity in the lens allows the ultrasound traveling along the shorter central path to be delayed so that the ultrasound along the longer marginal paths can catch up, aligning the compressions and decompressions to provide the highest voltage excursions in the piezoelectric transducer element for ultrasound waves originating from the focal region. **(b)** Electronic delay lines in the signal path after the piezoelectric elements provide the same focusing function as the lens. Apodization *shading* of the sensitivity of the transducers at the edge of the aperture suppresses aperture diffraction side-lobes in the beam. The mechanical focus depth is determined at the time that the ultrasound transducer is manufactured; it cannot be altered by the examiner. The electronic focus depth can be adjusted to a different depth in a microsecond, the time that ultrasound travels 1.5 mm, so that the electronic focus can be adjusted to each depth during the time that the ultrasound is traversing the tissue

Amazon EX1014 Page 859

vided in the image thickness direction by a cylindrical silicone rubber lens and in the lateral image direction by adjusting the relative phase of the adjacent transducers (Fig. 21.12b).

21.3.6 Measurement of Ultrasound Wavelength

Modern diagnostic ultrasound systems use sophisticated transducers and filters to improve resolution by shifting and broadening the ultrasound frequency band used for imaging, Doppler and other methods. It is possible, however, to use basic wave physics to measure the effective ultrasound wavelength of the system. In addition to refraction, ultrasound waves are also subject to diffraction. By placing a diffraction grating in the ultrasound image path in water, and imaging a point reflector, the diffraction pattern of the wave can be imaged and the wavelengths computed (Figs. 21.13, 21.14).



21.3.7 Attenuation of Ultrasound

As an ultrasound passes through tissue as a pulse or as a *continuous wave* along the beam pattern, some of the ultrasound energy is converted to heat in the tissue and some is reflected or scattered out of the beam pattern. In addition, nonlinear propagation converts some of the energy to higher harmonics ($2f$, $3f$ ...) of the fundamental frequency (f) of the wave. Attenuation increases with frequency, and the harmonic frequencies also may be outside the bandwidth of the detection system. These factors contribute both to actual and apparent attenuation of the ultrasound energy in the pulse and the average power passing through a cross section of the ultrasound beam as a function of depth. Changes in intensity are related to these changes in power, but the intensity is power divided by the cross-sectional area of the beam, so changes in beam cross-sectional area cause intensity changes that are not related to attenuation. Focusing the beam pattern can decrease the area, increasing intensity; diffraction and refraction increase the area, decreasing intensity.

Ultrasound attenuation is usually expressed as a ratio of the output power divided by the input power. In linear systems, this ratio is constant for all ultrasound intensities. In nonlinear systems, the relationship is more complicated. The attenuation rate of ultrasound in tissue

Fig. 21.13a–d Grating diffraction. Like any wave, an ultrasound beam can be tilted by a diffraction grating. The angle of tilt can be computed using Huygens' principle. If the pitch spacing of the grating is P, and the wavelength is λ, then the angle of deflection of the wave is expressed as $\sin\varphi = P/\lambda$. An ultrasound transducer array can act as a diffraction grating. This should be avoided. A diffraction grating in a water tank can be used to measure the wavelength of ultrasound. An image of a reflector deeper than the grating appears three times in the image, once at the true location and once for each of the diffraction orders. **(a)** Plane wave diffracted by grating. **(b)** Computaion of diffraction angle. **(c)** Ultrasound scanhead, diffraction grating and target. **(d)** Expected ultrasound image for a single frequency transducer

Amazon EX1014 Page 860

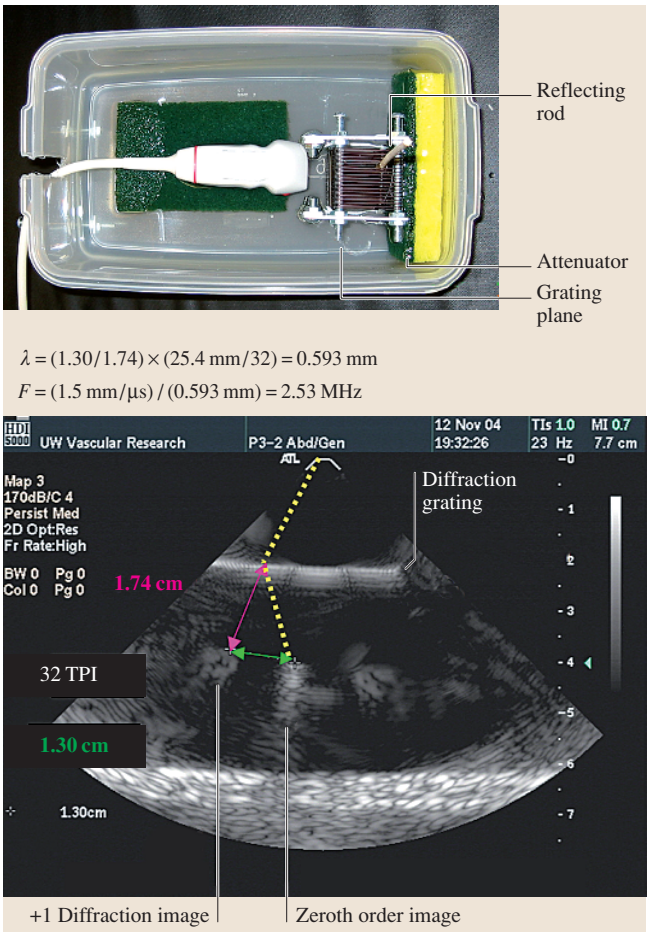


Fig. 21.14 Measuring ultrasound frequency and bandwidth by viewing a reflecting rod through a diffraction grating. The cross section of a rod forms the zeroth-order spot in the image. In addition to passing the ultrasound beam pattern directly (the zeroth image) the diffraction grating forms a +1 branch of the beam pattern (shown in *yellow*), which makes an image of the rod to the left (+1 diffraction image) and a symmetrical −1 image on the right. By measuring the geometric sine of the deflection angle, (1.30/1.74), the ultrasound wavelength can be computed using the spacing of the grating. The grating is made of monofilament fishing line wound on a frame made of 10–32 threaded screws (32 threads per inch, pitch spacing = 0.8 mm)

can be computed using base 10 (dB/cm, for engineers), base "e" (Np/cm, for scientists), compound percentage (% per cm, for physical therapists) or base 2 (half-value layers, for systems designers). The decibel, derived from the bel can be traced to Alexander Graham Bell, and the neper to John Napier. The bel is a factor of 10 in power, the neper is a factor of e (2.718) in amplitude. Ultrasound attenuation in soft tissue is between 3 dB/(MHz cm) and 0.3 dB/(MHz cm). Using intensity as the dependent variable in attenuation computations is complicated by changes in beam area. Some authors apply the bel to intensity. In acoustics the pressure amplitude of the wave, which is proportional to the square root of intensity (assuming constant impedance) is often used. The base 2 system is used to identify the thickness of tissue that attenuates the power by half, the intensity by half or the amplitude by half. The thickness is called the *half-value layer*, often without qualifying whether it is *half power*, *half intensity* or *half amplitude*. Intensity is proportional to the square of the amplitude, which is a multiplier of 2 in logarithms. If the cross-sectional area of the ultrasound beam is constant

$$\begin{aligned} 0.7\ \text{nepers} &= 6\ \text{dB} \\ &= 1\ \text{half-amplitude layer} \\ &= 2\ \text{half-power layers}\ . \end{aligned} \qquad (21.10)$$

Often the attenuation rate is nearly proportional to frequency, and frequency is inversely proportional to wavelength, so the attenuation can be most easily expressed in wavelength (Table 21.4). The use of half-value layers has the advantage that it can be related directly to the number of bits in a digitizer, simplifying computation.

Because of the high instantaneous intensities used in pulse-echo diagnostic ultrasound, all of the ultrasound propagation is nonlinear. The commonly used linear equations do not apply. The beam patterns are

Table 21.4 Ultrasound attenuation

Tissue type	Half amplitude (quarter power) thickness in wavelengths	60 dB pulse-echo attenuation depth in wavelengths (10 bits dynamic range)	Attenuation dB/(cm MHz)
Bone	1	5	30
Cartilage	8	40	10
Tendon	8	40	10
Skin	25	125	3
Muscle	25	125	3
Fat	80	400	1
Liver	80	400	1
Blood	300	1500	0.25

not as predicted, because the wave propagation speed in the higher-pressure regions is faster than in the lower-pressure regions. The waves convert from sinusoidal to sawtooth, introducing a series of harmonic frequencies into the wave. The change in frequency (wavelength) affects the ultrasound beam pattern. The conversion of ultrasound energy into harmonic frequencies also increases the attenuation because the attenuation increases with frequency.

A significant development was the use of tissue-generated harmonic signals for diagnostic imaging. The utility of these harmonics was discovered by accident when investigators were searching for better methods of detecting ultrasound contrast agents. Ultrasound contrast agents consist of microbubbles that have natural oscillation frequencies near diagnostic frequencies. The bubbles therefore oscillate in response to incident ultrasound. The bubble oscillation stores some energy, which is reradiated at the original ultrasound frequency as well as at harmonics of that frequency. Older ultrasound systems were, by nature, insensitive to the second harmonic of the natural transducer frequency. By using a transducer with an increased bandwidth (3 MHz center frequency transducer manufactured with the quality factor reduced spreading the bandwidth to 1.9–4.1 MHz) then lowering the transmit frequency (2 MHz) and raising the receive frequency (4 MHz), a system sensitive to the harmonic emissions from the contrast bubbles can be achieved. Surprisingly, harmonic echoes are generated by tissues even when contrast bubbles are not present.

21.4 Methods of Medical Ultrasound Examination

Over the half century of ultrasonic medical examination, a variety of examination methods have been tried, discarded, with some to be resurrected and honed for use.

A passive method was evaluated to detect local ultrasound emissions above background due to local temperature elevations of cancerous tumors. The emissions of about 10^{-13} W/cm^2 (1 nW/m^2) are present in all warm tissues and provide a minimum intensity for the operation of active ultrasound systems. Transmission ultrasound systems have been developed using an ultrasound transmitter on one side of a body part and a receiver on the other side to detect differences in attenuation (like X-ray imaging). These systems determine ultrasound attenuation and ultrasound speed through the intervening tissues. One version of this system used an acousto-optical holographic system to form the image. Several systems have been explored to insonify tissue from one direction and gather the scattered ultrasound from another direction. Such systems usually require an array of receiving transducers.

The only ultrasound systems that have gained wide acceptance in clinical use are: (1) the continuous-wave (CW) backscatter Doppler system, and (2) common-axis pulse-echo backscatter imaging and Doppler systems. The reason for the failure of the other systems is due to two features of medical ultrasound examination: refraction of ultrasound in tissue and the cost of receiving and processing data from array receivers. The CW methods are used only for Doppler measurements. The first of these systems was constructed before 1960 [21.45, 46]; their application in medical diagnosis was well established by 1970 [21.47–52]. Pulse-echo methods have been under continuous development since 1950. The majority of medical diagnostic imaging examinations in the world are performed with pulse-echo instruments.

21.4.1 Continuous-Wave Doppler Systems

Continuous-wave Doppler was used in one of the earliest ultrasound systems to be commercialized. Developed

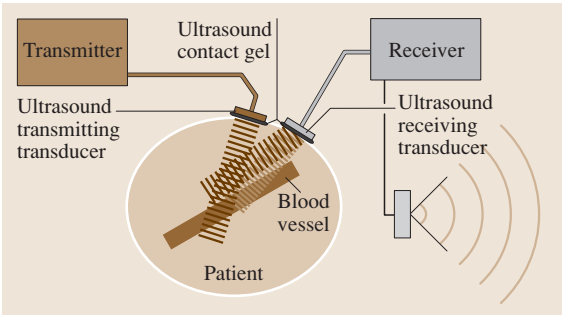


Fig. 21.15 Nondirectional continuous-wave Doppler. In a typical CW 5 MHz nondirectional Doppler, the incident ultrasound beam is scattered by stationary tissue forming a 5 MHz echo and by moving blood forming a 5.0004 MHz echo. The two echoes arrive at the receiving transducer simultaneously. By constructive and destructive interference between these waves, the received 5 MHz signal is amplitude modulated by the 5.0004 MHz echo. The frequency of the amplitude modulation is the difference frequency 5.0004 – 5 MHz or 400 Hz, which is amplitude demodulated, providing a 400 Hz audio signal to the speaker

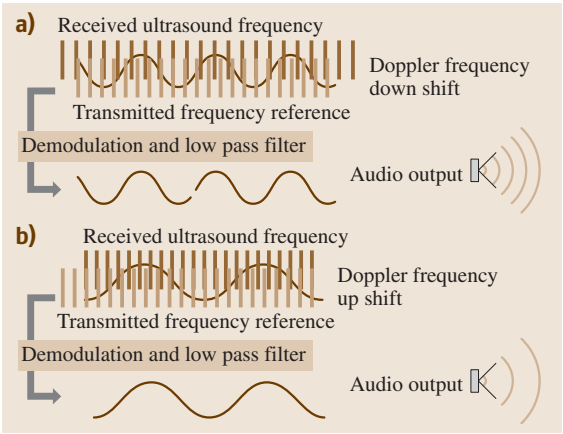


Fig. 21.16a,b Interference between echoes to produce audio frequencies. Constructive interference between the Doppler frequency-shifted echo and the original frequency wave increases the resultant amplitude; destructive interference reduces the amplitude. This occurs whether the Doppler shift is downward **(a)** due to reflections from blood cells receding from the transducer, or is upward **(b)** due to reflections from blood cells advancing toward the transducer. By amplitude demodulating the result, an audio frequency signal results. Amplitude demodulation can be done by multiplying the two ultrasound frequency signals together and low-pass filtering the result or by squaring (rectifying) the sum of the two ultrasound signals and low-pass filtering the result

independently in Japan in 1957 (Satamura and Koneko), and in the US in 1959 (Baker et al.), a system can be constructed from an RF transmitter, transmitting piezoelectric transducer, receiving piezoelectric transducer and amplitude demodulating receiver (Fig. 21.15). In the region of tissue where the transmitting beam pattern overlaps with the receiving beam pattern, the reflected wave from moving tissue will combine to interfere with the reflected wave from stationary tissue. The reflected wave from moving tissue has a Doppler shift. The Doppler-shifted frequency will interfere with the reflected wave from stationary tissue to cause an amplitude modulation in the combined wave. The frequency of that modulation is within the hearing range.

Ultrasonic Doppler examination produced an audible frequency by chance. The best ultrasound frequency to reflect by Rayleigh scattering from blood at a depth of 2 cm is 5 MHz. This frequency is shifted by the ratio of twice the blood speed ($2 \times 0.75\ \text{m/s} = 1.5\ \text{m/s}$) to the ultrasound speed 1500 m/s ($1.5/1500 = 0.001$) to cause a downshift in the ultrasound frequency of 5 kHz. The interference between the unshifted and shifted waves causes a 5 kHz modulation in the amplitude of the 5 MHz wave. This modulation is converted to sound by common amplitude modulation (AM) radio circuitry.

The introduction of the transistor in 1960 allowed this simple circuit to be produced in a palm-size package so that convenient clinical examination could be done. Whether the blood is approaching the transducer or receding from the transducer (Fig. 21.16), the audio signal is the same, with audio frequency proportional to the blood velocity.

These CW Doppler systems allowed the detection of narrow arterial stenosis by the high velocities present in the stenotic lumen. Normal velocities are below 125 cm/s, causing a Doppler shift of 4 kHz if the Doppler angle is 60°. The blood velocity in a stenosis is limited by the blood pressure via the Bernoulli effect; the velocity can be no greater than 500 cm/s for a systolic blood pressure of 100 mmHg. A blood velocity of 500 cm/s causes a Doppler shift of 16 kHz. These CW Doppler systems also allow the detection of venous obstruction by the loss of variations in venous velocity synchronized with respiration, and the testing of venous valve incompetence by the detection of reverse flow resulting from distal compression release or sudden proximal compression. An examiner with one year of

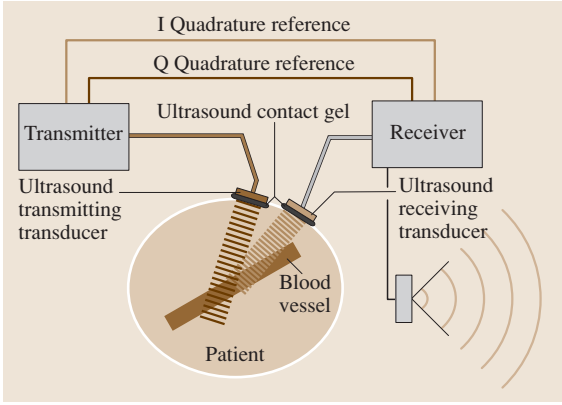


Fig. 21.17 Directional continuous-wave Doppler. By using a pair of reference waves at the transmitted ultrasound frequency, but phase shifted with respect to each other, the receiver Doppler demodulation can be performed against each of the reference waves. This allows the instrument to differentiate Doppler shifted echoes with increased frequency due to blood cells approaching the transducer from those with decreased frequency due to blood cells receding from the transducer

Amazon EX1014 Page 863

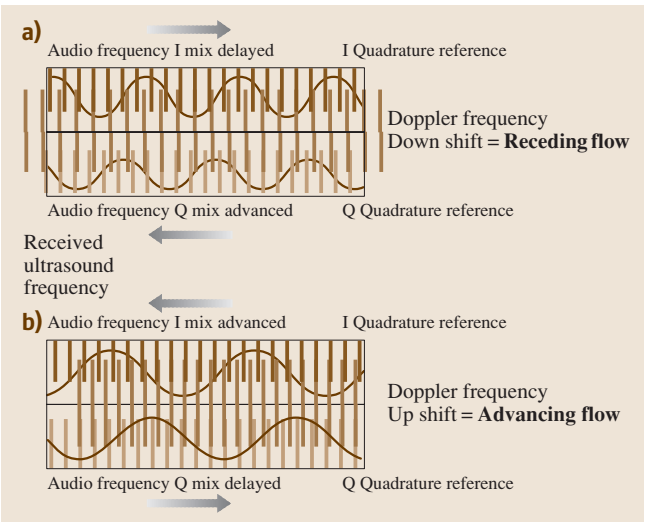


Fig. 21.18a,b Quadrature Doppler demodulation. **(a)** Demodulation of the of a Doppler down shifted echo against an-I reference and Q reference produces a quarter-wave advanced signal (shifted to the right) on the I product compared to the Q product. **(b)** Demodulation of the of a Doppler upshifted echo against an I reference and Q reference produces a quarter wave delayed signal (shifted to the left) on the I product compared to the Q product

training and experience can diagnose all superficial arterial obstructions and venous obstructions and reflux with accuracies exceeding 95% against any gold standard.

The addition of directional detection by Fran McLeod in 1968 simplified the separation of mixed arterial and venous flow, and allowed more-objective evaluation of velocities (Fig. 21.17). Directional detection was accomplished by providing a pair of quadrature signals derived from the transmitted frequency directly to the demodulator in the receiver. Quadrature signals differ only in phase by a quarter cycle. When the Doppler-shifted ultrasound echo is demodulated against each of the quadrature reference signals, the pair of difference waves produced are phase shifted by 90° (Fig. 21.18). If the reflecting blood is advancing toward the transducer then audio frequency (AF) I wave is advanced compared to the AF Q wave, if the reflecting blood is receding, then the AF I wave is delayed with respect to the Q wave. By further delaying the AF Q wave by a quarter cycle (the shift time is dependent on frequency), and adding or subtracting the resultant AF I and AF Q delayed, the sound channels can be separated (Fig. 21.19). After separation, the Doppler signal representing advancing flow can be delivered to one ear and the Doppler signal representing receding flow can be delivered to the other ear. Although these continuous-wave Doppler systems are simple and have not changed substantially since 1970, they still have tremendous clinical utility and are used by angiologists and vascular surgeons in daily practice. A CW Doppler transducer is provided with modern echo-cardiology systems (Fig. 21.20) for the measurement of cardiac output and for accurate velocity measurements of blood flow jets associated with valve disease.

During 1970–1985, several arterial mapping angiographs (Fig. 21.21) were developed to associate high-velocity Doppler signals, indicating stenoses with arterial landmarks such as bifurcations. These angiographs used mechanical or other methods to track the location of the ultrasound beam and display the results on an image that could be compared with an X-ray angiogram. Modern, large-field-of-view three-dimensional ultrasound imaging systems use extensions of this method.

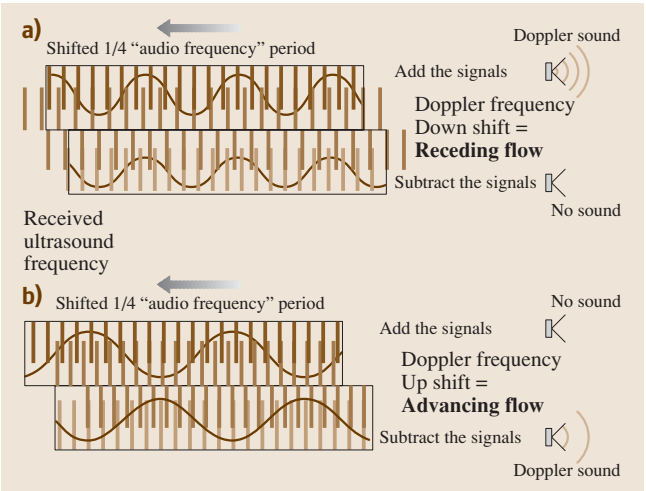


Fig. 21.19a,b Audio direction separation. By retarding the I quadrature channel by 1/4 of the period of the audio frequency, and then adding or subtracting the resultant pairs, the directions of flow can be separated. **(a)** The added pair results in sound for only the downshifted Doppler signals representing receding flow. **(b)** The subtracted pair results in sound for only the upshifted Doppler signals representing advancing flow

Amazon EX1014 Page 864

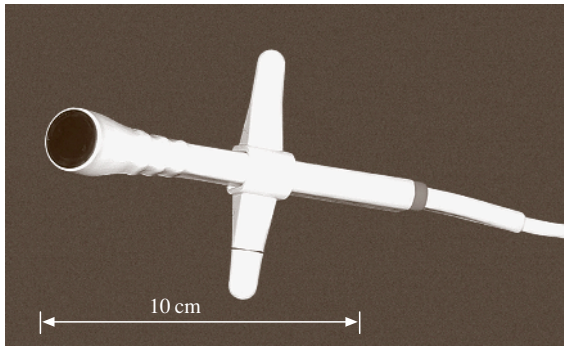


Fig. 21.20 Dual-element continuous-wave Doppler transducer for cardiac output measurements. This transducer is designed to fit in the suprasternal notch between the clavicles at the base of the neck. From this location, the Doppler ultrasound beam patterns can be nearly aligned with the axes of the cardiac valves. The ultrasound avoids the lungs by passing directly down through the mediastinum

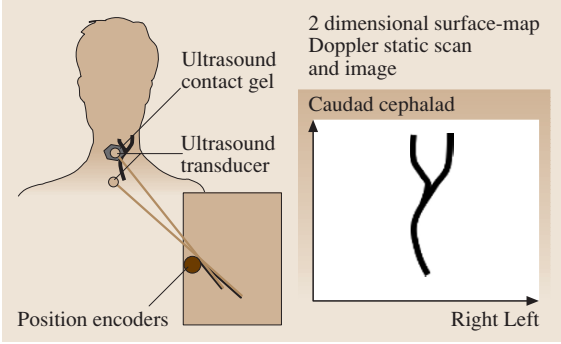


Fig. 21.21 Doppler arteriography. By monitoring the position of a Doppler ultrasound transducer, transferring the position on a storage image screen, and marking the display screen only when the Doppler detects blood flow, an image of the flow can be generated. Most of these systems used continuous-wave Doppler; the Hokanson system used pulsed Doppler

21.4.2 Pulse-Echo Backscatter Systems

Pulse-echo ultrasound consists of sending a pulse of ultrasound into tissue and receiving the echoes from various depths. The length of the transmitted pulse may be as short as 0.5 μs (for B-mode imaging) or as long as 5 μs (for trans-cranial Doppler). Usually the same transducer or transducer aperture is used for both transmitting and receiving. The depth of each reflector is determined by the time of flight between the transmitted pulse and the received echo: a reflector 3 cm deep returns an echo at 40 μs; a reflector 15 cm deep returns an echo at 200 μs. If the system is examining to a depth of 18 cm, then a new pulse can be transmitted after 240 μs. In that case, an echo returning at 40 μs after the later transmit pulse might come from 21 cm deep (280 μs) rather than 3 cm. Tissue attenuation will suppress the echo from deep tissue by about 54 dB (3 MHz ultrasound) compared to the echo from 3 cm.

Table 21.5 Ultrasound examination parameters

Maximum Examination depth (cm)	Ultrasound wavelength (mm)	Ultrasound frequency (MHz)	Pulse repetition Period (μs)	Pulse repetition Frequency (kHz)
3	0.075	20	40	25
5	0.125	12	67	15
6	0.15	10	80	12.5
10	0.250	6	133	7.5
12	0.3	5	160	6.25
15	0.375	4	200	5
20	0.5	3	266	3.75

To perform a pulse-echo examination, an ultrasound transducer (or array of transducers) is positioned to form a beam pattern along an axis, which is assumed to penetrate tissue in a single straight line. A short ultrasound burst is launched along that line. After allowing for the round-trip travel time to the depth of interest (Table 21.5), the ultrasound echo from a segment of tissue in the beam pattern is received, demodulated and prepared for display (Fig. 21.22). The ultrasound transmit pulse can be adjusted for frequency, duration (Fig. 21.23), energy, and aperture size and shape to fine-tune the examination if needed. The ultrasound echo can be demodulated to yield the amplitude (Fig. 21.24) and/or the phase. The results of multiple pulse-echo cycles along a single beam pattern can be combined. The echo is usually amplified with a fast-slew programmed amplifier to increase the gain of later echoes coming from deeper tissues, because they will be attenuated more than shallow echoes. If tissue visualization is desired, the echo strength is displayed in an image along a line representing the direction of the ultrasound beam, with the location along the line (representing depth) determined by the time after transmission that the echo returns. If a measurement of tissue motion is desired, then additional identical pulses are transmitted along the same beam pattern to obtain a series of nearly identical echoes from each depth of interest. These echoes are examined for changes in phase, which is the most sen-

Amazon EX1014 Page 865

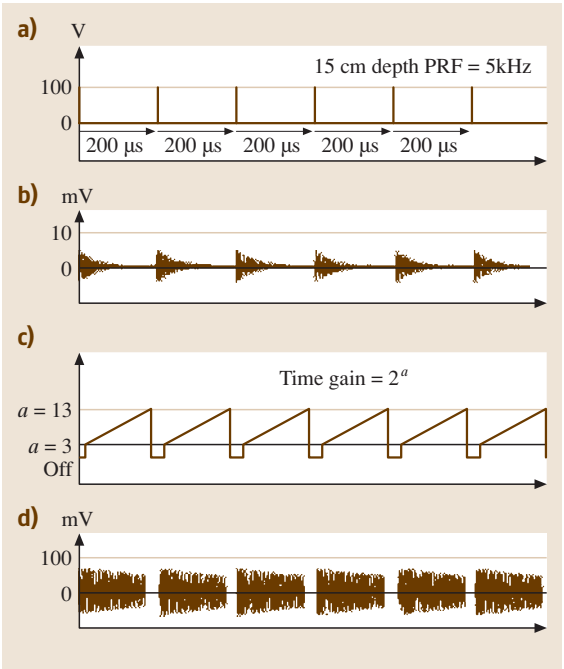


Fig. 21.22a–d Time gain compensation (TGC). **(a)** A new ultrasound pulse is transmitted each pulse repetition period (200 µs in this case). **(b)** Echoes from shallow and deep depths return at known times because of the speed of ultrasound and the expected round-trip travel time. The later echoes received from deeper tissues are weaker than the echoes received from shallow tissues because the echoes from deeper tissues have passed through more-attenuating tissue. **(c)** An amplifier programmed to increase the gain for deep echoes more than for shallow echoes is used to compensate for attenuation. This depth gain compensation (DGC) or time gain compensation (TGC) amplifier is under operator control. **(d)** The echoes, after TGC are ready for demodulation by analog or digital methods

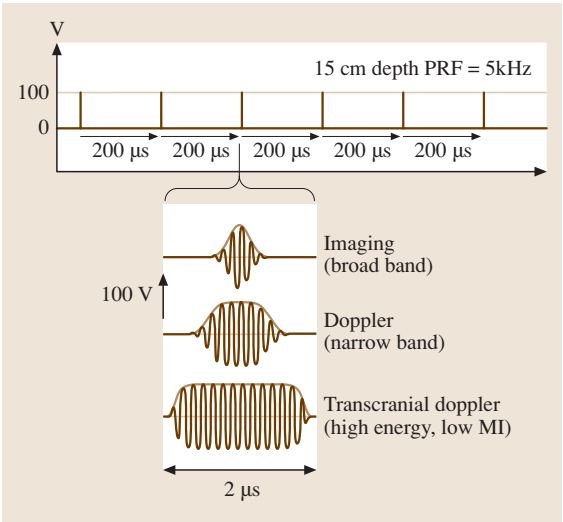


Fig. 21.23 Different ultrasound transmit bursts for imaging and Doppler applications. Short broadband transmit bursts are used for B-mode imaging; long narrow-band transmit bursts are used for Doppler applications. A longer transmit burst contains more energy if the maximum acoustic negative pressure (mechanical index) are limiting factors. A short transmit burst allows better depth resolution and also spreads the diffraction side-lobes of the beam pattern, suppressing lateral ambiguity. A longer transmit burst is preferable for Doppler applications where echoes from blood are weak and coherence length is important for demodulation

sitive indicator of displacement of the tissue reflecting the ultrasound.

This simple model places constraints on the ultrasound examination. First, at any depth, the strongest echoes from a bone or air interface perpendicular to the beam are 10^6 (60 dB) stronger than the weakest echoes from cells in liquid (blood). So, a 10 bit digital echo-processing system is the minimum to handle the range of echo strengths expected in an image. Second, because the speed of ultrasound is about 1.5 mm/µs in most tissues, echoes returning from 15 cm deep arrive 200 µs after ultrasound pulse transmission. A new pulse cannot be sent into the tissue along the same line or along another line until all of the echoes from the first pulse have returned. So, the maximum pulse repetition frequency (PRF) for an examination to a depth of 15 cm is 5 kHz (1/200 µs). Ultrasound energy is attenuated to 1/4 of the incident energy in most body tissues after passing through a depth of 80 wavelengths. So after passing through a depth of 400 wavelengths and returning through 400 wavelengths, the echo strength is only 0.25^{10} or 10^{-6} of the strength of a similar echo from the shallowest tissue. To compensate for this attenuation, an amplifier must amplify the deep echoes 60 dB more than the shallow echoes (Fig. 21.22). An alternative is to digitize the echoes with a 20 bit digitizer, 10 bits for the dynamic range of the echoes from different tissues and 10 bits to account for attenuation. In either case, the maximum depth of an examination is limited to 400 wavelengths of ultrasound. An ultrasound depth resolution cell (a pixel) is equal to about twice the wavelength of ultrasound. So, an ultrasound image can contain 200 pixels in the depth direction (Table 21.5). These con-

Amazon EX1014 Page 866

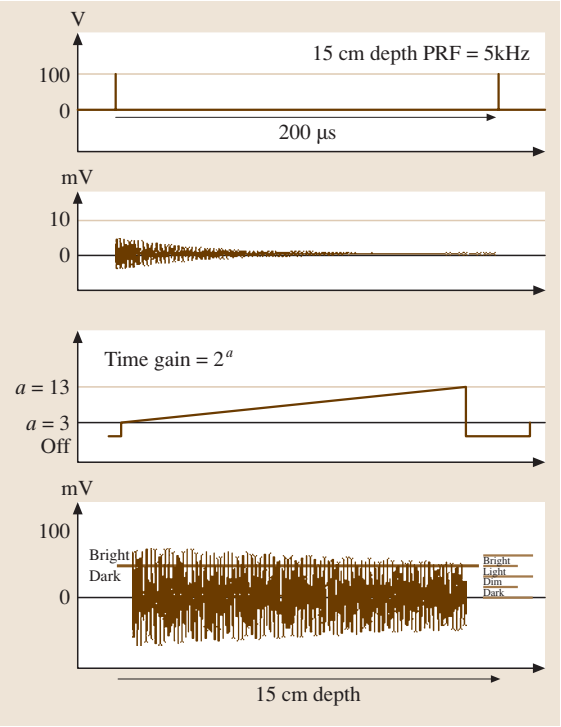


Fig. 21.24 Preparing the amplitude demodulated brightness mode echo for display. The amplitude of the amplified echo can be converted into brightness values by a digitizer. In early B-mode scanners, the display could show bright (echo) or dark (no echo) based on a one bit threshold detector (*dark line*). A two-bit digitizer allows four levels of gray scale (scale on *right*), making the result less sensitive to the setting of the TGC amplifier. Modern ultrasound B-mode imaging systems use an eight-bit digitizer to display 256 levels of gray. Depending on the method of presentation, the eye can distinguish no more than 64 stepped levels of gray or 16 isolated levels of gray

siderations provide a practical relationship between the ultrasound frequency and the PRF, based on the attenuation of tissue. The PRF is 1/800 of the ultrasound frequency in most applications.

21.4.3 B-mode Imaging Instruments

A series of B-mode (brightness mode) ultrasound imaging methods have been developed over the last 40 years. The methods have used the latest electronic technologies to improve the speed, convenience, capability and utility of the systems. The original imaging systems were based on radar technology and analog electronic

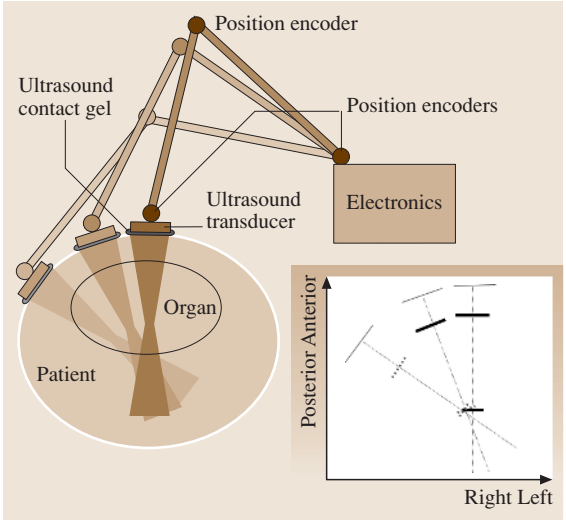


Fig. 21.25 B-mode static scanner. The traditional B-scanner restricted the transducer position and direction to movements in one plane that would form a cross-sectional image through the patient. The position of the transducer was located on the image display, and a line projected from the origin along the display in the direction of sound beam propagation. During the scan a marker was moved along the ultrasound beam direction at a speed of 0.75 mm/µs, representing the time that an echo would return after a round trip to each depth when traveling at 1.5 mm/µs. If a strong echo was detected, then a mark was stored on the screen to contribute to the image. The transducer was scanned across the skin to form an ultrasound scan cross-sectional image of the tissue

parts. Today, ultrasound systems are personal computers using standard operating systems with specialized display software coupled to custom ultrasound circuitry on the front end. A review of the history is useful because the modern instruments keep incorporating older methods.

Two-Dimensional Static B-mode

The original B-scanners (Figs. 21.25, 21.26) consisted of a single ultrasound transducer element coupled to a mechanical system to track the origin and direction of the ultrasound beam. The systems had the advantage that large fields of view could be scanned, such as a plane through a full-term fetus, and the sonographer could return to portions of the image to *paint in* anatomical features by approaching each of the reflecting interfaces from a perpendicular direction. As the image display system showed the brightest echo for

Amazon EX1014 Page 867

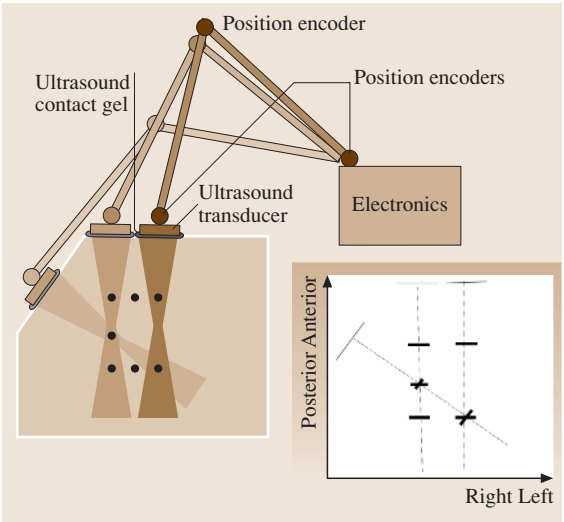


Fig. 21.26 B-mode calibration phantom. Analog electronics were not stable at the time that static B-scanners were constructed. Calibration phantoms were used to assure that a target would appear in the same location when viewed from different locations. Adjustments were provided to align the electronics

each anatomic location, *compound scanning* was used to obtain echoes from a variety of directions at each tissue location to render a useful image. Since a scan might take 5 min, movement of the patient during the scanning period was a problem, rendering the image useless. Because these systems used analog electronic

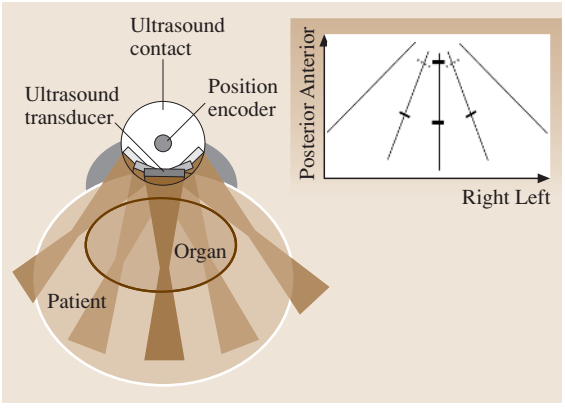


Fig. 21.27 Real-time two-dimensional ultrasound images. Motor-driven ultrasound transducers were able to rapidly tilt to a series of angles obtaining data from a series of lines in a plane of tissue to form a new two-dimensional image 10 times or 30 times per second

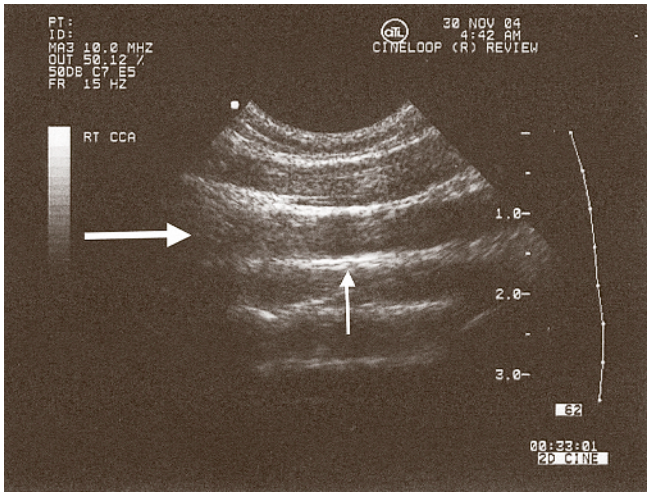


Fig. 21.28 Two-dimensional B-mode (brightness mode) common carotid artery image. The lumen of the common carotid artery in longitudinal section appears as a darker streak (*wide arrow*) bounded by the superficial and deep walls separated by 6 mm. The walls appear brighter where the ultrasound scan lines are perpendicular to the walls (*narrow arrow*) because of the specular backscatter.

devices, which were subject to deterioration over time, the stability and alignment of the instruments was important. In addition, since the same tissue was examined from multiple locations, the co-registered alignment of the tissue voxels (small volumes of tissue) on the image

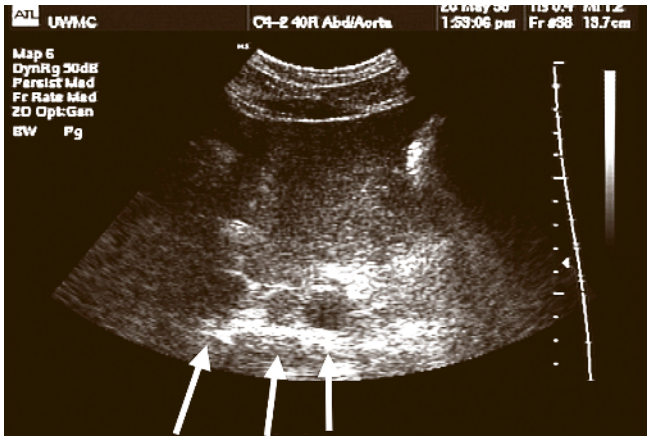


Fig. 21.29 Duplication of aorta [21.53] in the image. Refractive tissues composed of fat ($c = 1.45\,\mathrm{mm/\mu s}$) and muscle ($c = 1.58\,\mathrm{mm/\mu s}$) at the midline of the abdomen form two prisms. Each prism allows the aorta to be viewed resulting in two images of the aorta. This effect has also caused double viewing of a single pregnancy

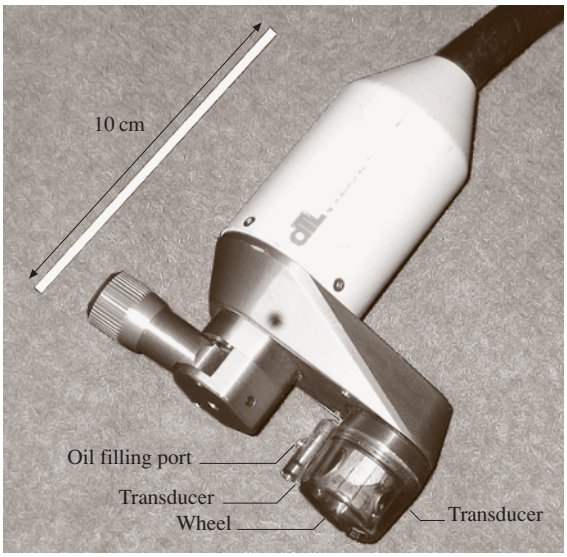


Fig. 21.30 Mechanical scan head. Three transducers are mounted around a rotating wheel encased in a cover filled with oil. Two of the 3 transducers can be seen in the image. Between the transducers is an air bubble in the oil; the air bubble must be removed through the oil filling port before scanning. Under the bubble is a rectangular shape that is the housing for a switch. Each of the transducers has a switch mounted opposite the transducer on the wheel which is activated magnetically to connect the transducer when the transducer is pointed toward the patient. The transducer wheel is driven by a motor in the white housing through a belt. A control knob on a joystick allows adjustment of the Doppler ultrasound beam, when needed; swinging the arm tilts the beam, rotating the knob advances the sample volume along the line. Electrical coupling to the transducer was accomplished by using a transformer aligned on the axle of the wheel with one coil stationary and one rotating. The magnet-activated switches mentioned above selected the proper transducer that was pointing at the patient

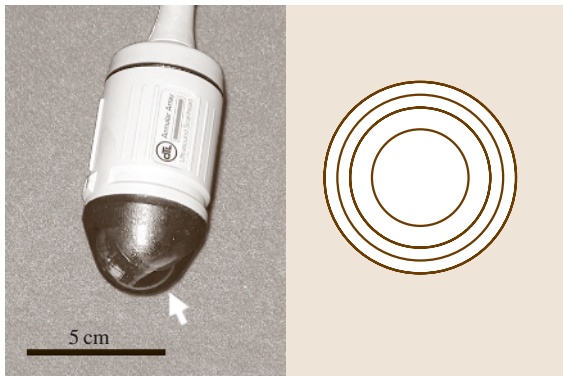


Fig. 21.31 Mechanically steered annular array scan head. The circular transducer is divided into concentric circles with each segment having the same area. By delaying by different amounts, the signals to the more central transducers, the depth of focus can be adjusted. A 10 ns delay between the transducer rings with the central transducer operating last will retract the focus a centimeter. The oil between the transducer and the dome shaped cap provides refractive focus in addition to the focus provided by the delays to the transducer. To point the ultrasound beam along different lines, the transducer wobbled back and forth rather than rotated. A rotating motion was not possible because the four coaxial cables could not cross the rotating axis. The arrow is pointing at the transducer along its axis

was critical. Regular testing of the instrument on alignment phantoms was required. The speed of ultrasound in the phantoms was controlled by the chemical composition of the phantom at the time of fabrication and the temperature of the phantom at the time of use. Of course, because the speed of ultrasound is not the same in different tissues, the success of the phantom in controlling image distortions in the depth direction during actual examination was limited. In addition, because the ultrasound beam can bend due to refraction as it passes from tissue to tissue, a factor ignored in the phantom, preventing lateral image distortions was impossible. However the phantoms did serve to prevent huge errors. Modern real-time ultrasound scanners that complete a scan in 30 ms (Fig. 21.27) form each image without overlapping beams so object co-registration is not tested. Thus, the value of phantoms has diminished. They still retain educational value.

Two-Dimensional Real-Time B-mode Mechanical Scanning

The earliest real-time systems scanned the ultrasound beam by tilting the ultrasound transducer, tracking the position and direction of the ultrasound beam with timing or an optical encoding system. The images had tremendous utility (Fig. 21.28) in spite of the poor lateral resolution compared to the depth resolution. Unfortunately, neither phantoms nor digital tracking of the ultrasound beam direction avoid image distortions due to refraction of the ultrasound beam in tissue (Fig. 21.29). These distortions, which are still a problem in the lateral direction, are ignored.

The ultrasound scan heads were convenient, ergonomic and compact (Fig. 21.30), located at the end of a cable attached to the ultrasound scanner. A variety

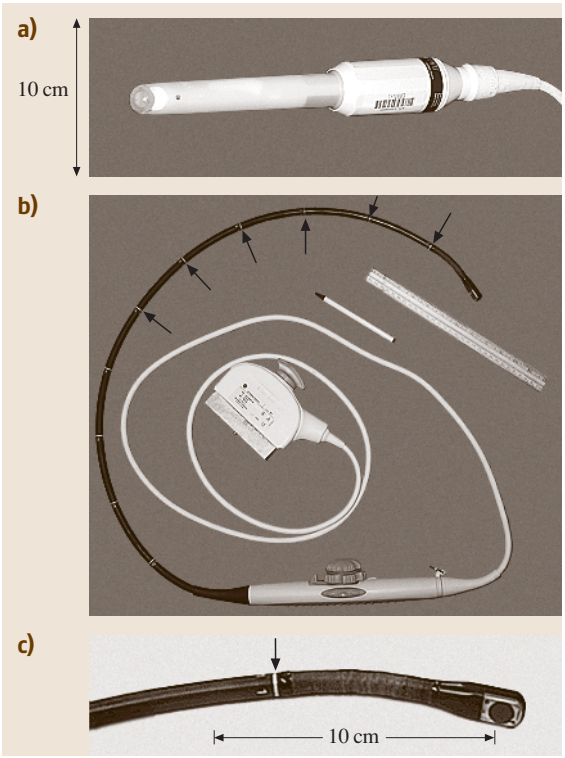


Fig. 21.32a–c Intracavity scan heads. **(a)** Intracavity *wobbler* scan head for transvaginal or transrectal examination. To provide acoustic coupling between the transducer and the housing, the scan head is filled with liquid. The transducer, at the left end of the image, is pointing out of the page. **(b)** Trans-esophageal (TEE) ultrasound scan head. The patient swallows the end with the transducer array (*upper right* in the picture) mounted in the end of the flexible esophageal probe. Marks every 10 cm (*arrows*) allow the cardiologist to monitor the location of the transducer array from the patient's teeth. **(c)** The trans-esophageal ultrasound transducer 5 MHz linear phased array has a circular boarder so that the array can be rotated to rotate the scan plane once the transducer is in the esophagus. A pair of control handles ((**b**) at *bottom*) allows the cardiologist to rotate the transducer array and to tilt the tip housing the array to examine the heart in different planes. Some patients tolerate this procedure well with little anesthesia, others require more sedation

of scan heads were developed including some with electronic focusing (Fig. 21.31) and specialized shapes for placement in the esophagus, rectum, vagina (Fig. 21.32), and some 1 mm in diameter for insertion inside blood vessels.

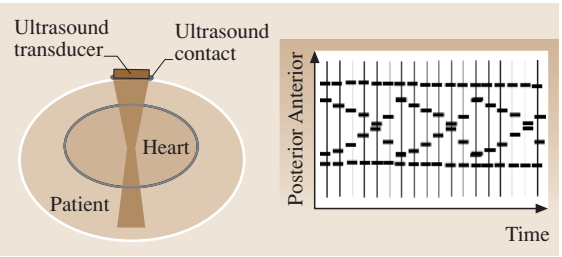


Fig. 21.33 M-mode (motion-mode) examination. For the study of cardiovascular wall motion, a single ultrasound beam is directed through the heart or other tissue. A *horizontal mark* showing the depth of each reflecting tissue in the vertical direction is shown on each vertical image line. By transmitting a new ultrasound pulse every millisecond, the echoes from tissue interfaces can be plotted for each pulse with time on the *horizontal axis*. By placing the lines side by side, the trajectories of tissue motion along the ultrasound beam are displayed. Speed of the reflecting tissue interface along the direction of the ultrasound beam is shown as a tilt or slope of an interface displayed on the image

M-mode Ultrasound

Among the early applications of ultrasound was the study of the motions of the heart using M-mode (motion mode) imaging [21.54, 55] (Fig. 21.33). By holding the ultrasound beam stationary and plotting the image width as a function of time, motions of the heart and other vascular structures could be documented. The velocity of a valve could be determined (Fig. 21.34) by measuring the slope of the trajectory on the M-mode tracing.

Two-Dimensional Real-Time B-mode Electronic Scanning

An alternative to moving the ultrasound transducer to point in different directions, is to have an array of transducers [21.56]. By switching to each transducer in turn to acquire echoes from the corresponding beam pattern, a two-dimensional (2-D) image is formed (Fig. 21.35). Several ultrasound scanners developed between 1970 and 1980 used low-density arrays like this. A scanhead with more transducer elements, spaced closely together (Fig. 21.36), could use groups of transducers to form an aperture. These high-density arrays permit phase focusing to a series of depths (Figs. 21.36–21.38), phase tilting of the ultrasound beam patterns (Fig. 21.39) and apodization of the aperture to smooth the ultrasound beam pattern (Fig. 21.40). The transducer array is usually formed from a single rectangular piezoelectric crystal, cut into a row of thin rectangles to form

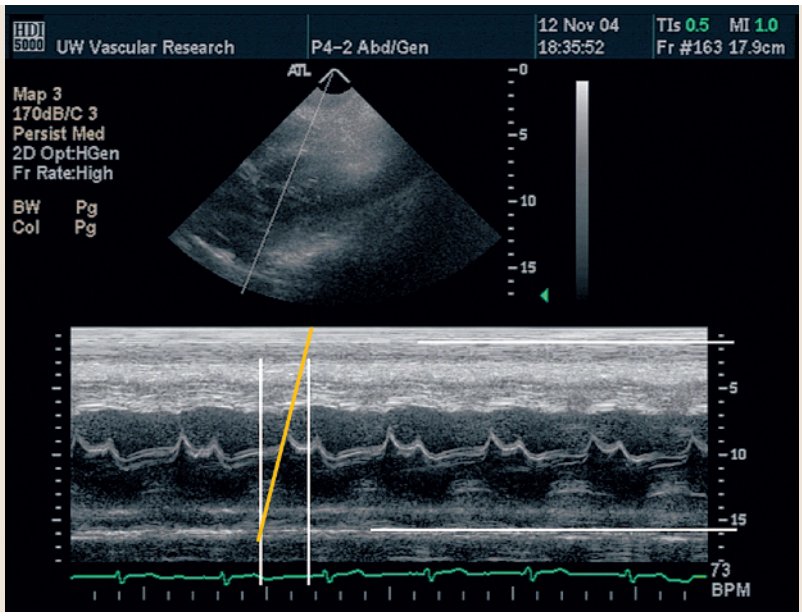


Fig. 21.34 M-mode (motion mode). Motion of a tissue interface along the ultrasound beam line is shown as depth (*vertical*) versus time (*horizontal*). The speed of the heart structure can be computed from the slope of the tracing. In this example, the speed is $158 - 16\,\text{mm}/0.4\,\text{s} = 355\,\text{mm/s}$. The green line across the bottom of the image is the Electrocardiograph (ECG, EKG) timing signal

an array in one direction. Because of the cuts, the array becomes flexible in the long dimension and can be curved (Fig. 21.41) to form a sector scan. Each of the 128 elements is connected via an electrical impedance-matching network to a coaxial cable, which forms part of the scan-head cable running back to the scanner. The plug connecting the 128 coaxial connectors to the ultrasound scanner has separate contacts for all 128 cables. Behind the plug in the scanner are separate ultrasound transmitters and receivers for each transducer element. Often each of the receivers will have a separate digitizer to convert the TGC-amplified signal into digital form.

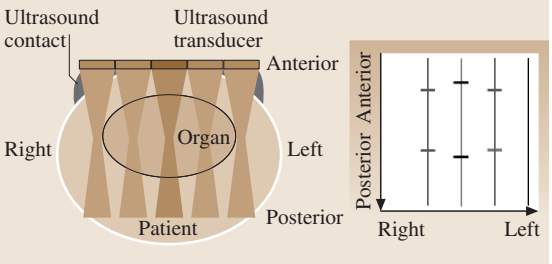


Fig. 21.35 Low-density linear-array two-dimensional B-mode. If the tissue does not move very much during the acquisition of an image, and the ultrasound beam pattern is moved across the tissue by successively activating different transducers and plotting the echoes with depth along the lines of the corresponding beam patterns, a two-dimensional image is formed. By convention in radiologic images, the images are usually displayed as if the patient is facing the physician. There is disagreement between clinics about how images should be oriented

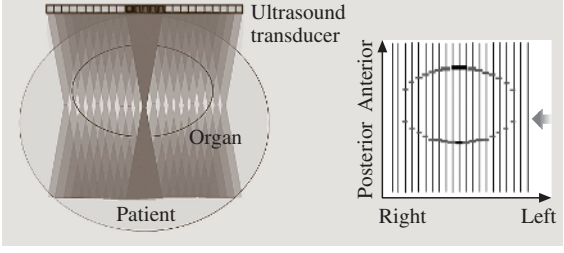


Fig. 21.36 High-density linear-array scan head. The transducers can be spaced more closely than the width of the ultrasound beam. Then each ultrasound aperture is formed by multiple transducers at the scan head face. The beam patterns overlap. The use of multiple transducers in the aperture allows the shaping of the aperture and electronic focusing of the ultrasound beam pattern

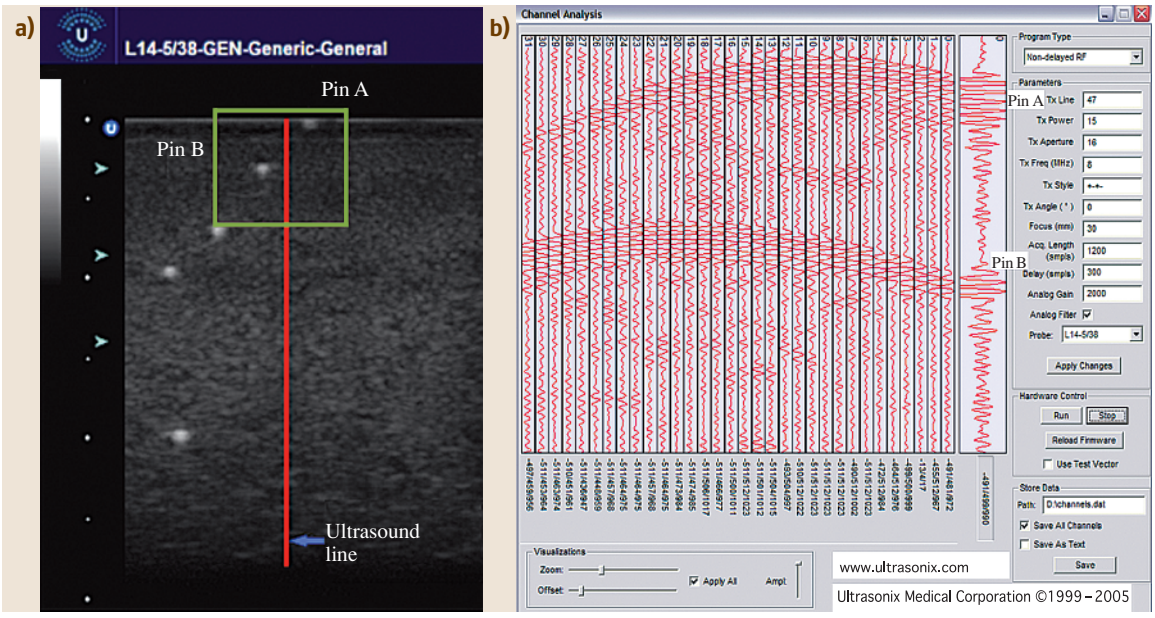


Fig. 21.37a,b Echoes received by each transducer in an array. By imaging highly reflective pins in a phantom, the effect of distance on travel time from a reflector to each transducer element in an array can be appreciated. By applying an appropriate delay on each of the transducer signals so that the echoes from the target align in time, and then adding the transducer signals together, a receiver beam pattern is formed, which is focused on the target. The time delays can be adjusted in 6 ns steps. **(a)** The transmit pulse was launched along the *red ultrasound line*. The echoes are displayed for data gathered from within the *green box*. **(b)** Radio-frequency echoes from each of 32 transducers are displayed for the echoes returning from a single transmit pulse (Courtesy Ultrasonix Medical Corp., Burnaby, www.ultrasonix.com)

The receivers and digitizers together form the ultrasound beam former for receive mode. Early array transducers had 32 elements, and the electronics were located in the scan head. Current instruments have 128 elements, a number limited by the costly cable containing 128 coaxial leads that connects the array to the receivers and transmitters in the console. The number of transducers will increase in the future as higher-capacity cables become available or, alternatively as it becomes possible to house the transmitting and receiving electronics in the ultrasound scan head.

A linear array transducer allows for adjustment of the transmit focus and aperture separate from the receive focus and aperture (Fig. 21.42). Because during the receiving period the relationship between depth and time is 0.75 mm/μs, a system can change receive focus

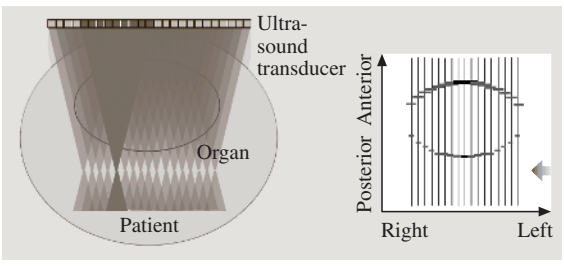


Fig. 21.38 Adjustable focus with large aperture. Using a high-density linear array to generate a large aperture and focus at a deep depth

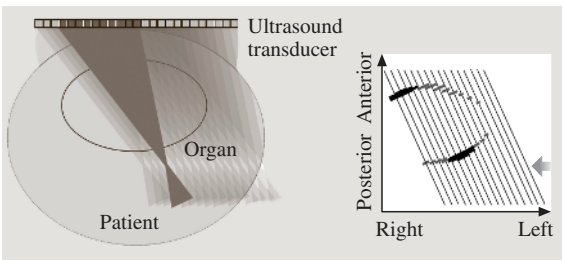


Fig. 21.39 Steering of the ultrasound beam pattern. Multiple transducers in the aperture allows the ultrasound beam to be steered at an angle for either imaging or Doppler by introducing phase delays in the signals to and from the transducers progressively along the transducer array

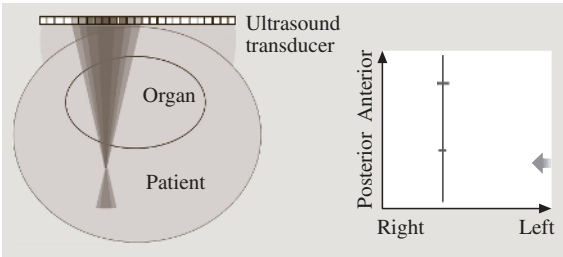


Fig. 21.40 Apodized aperture. Apodization is used to limit the side-lobes of the ultrasound beam pattern. Transducers near the edge of the aperture are excited with a low voltage during transmit compared to the central transducers and are amplified with a lower gain compared to the central transducers during receive. The drawing on the right shows the image line corresponding to the beam pattern with the superficial and deep echoes plus the mark on the right, indicating the focal depth

every 20 μs, advancing the focus 15 mm each time along with the corresponding aperture. This focal adjustment is called dynamic focusing and is similar in concept to time-gain compensation. It is also possible to acquire data along the ultrasound line using several pulse-echo cycles, each with a transmit burst that is focused at a different depth. Although dynamic receive focusing takes no more time than static focusing, using double transmit focusing to further improve the lateral resolution of the image at a greater range of depths does reduce the frame rate by half. The location of the transmit focus is indicated by a carrot on the right of the image.

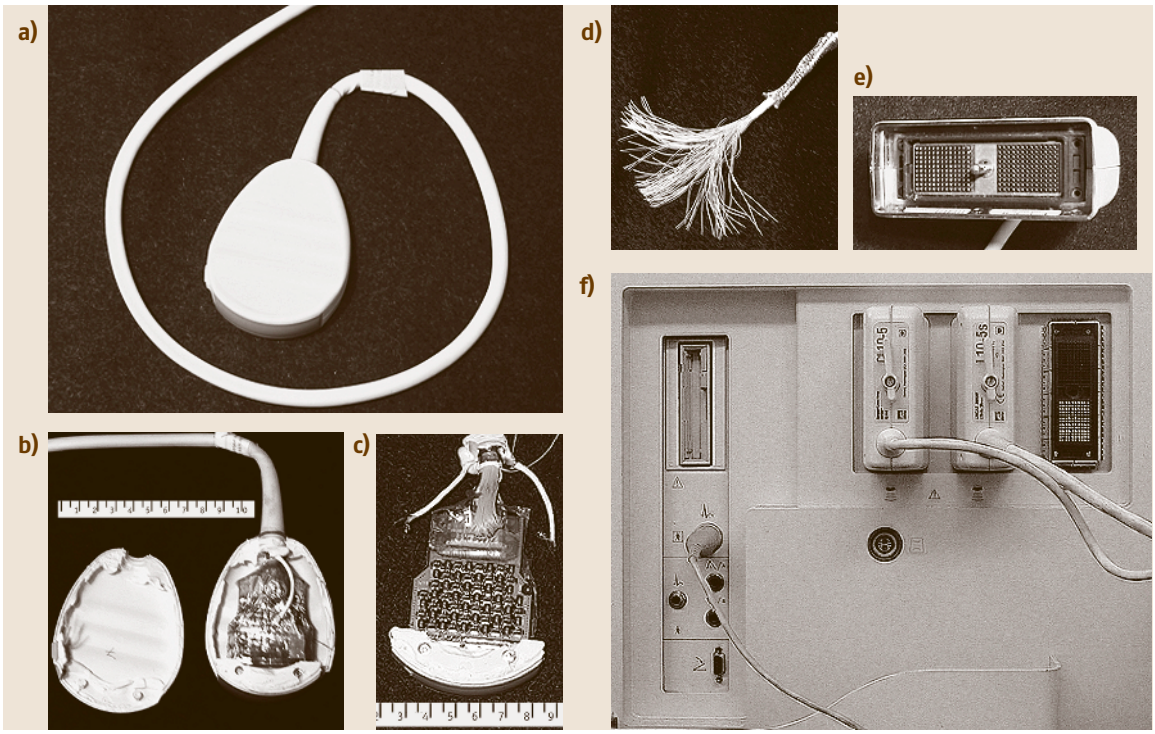


Fig. 21.41a–f Scan head details. This curved linear-array scan head consists of an array of 128 transducer elements cut from a single block of PZT piezoelectric ceramic. The array block is 64 mm long and 5 mm wide, cut into elements 0.4 mm long and 5 mm wide with a 0.1 mm kerf between. The block is curved to form a curved linear array. **(a)** Transducer in housing with cable; **(b)** Transducer housing removed showing the copper foil shielding over the internal electronics; **(c)** Shielding removed to show the 64 electronic matching circuits on this side of the circuit board; **(d)** 128 individual coaxial cables housed in the ultrasound scan head cable; **(e)** Scan head plug showing the connector electrodes for the 128 coaxial cables; **(f)** Front panel of the ultrasound scanner showing the sockets for the scan head connections. The front panel also shows connectors for other instruments including: electrocardiograph, plethysmograph, phonocardiograph microphone and computer disk drive for data exchange

Amazon EX1014 Page 873

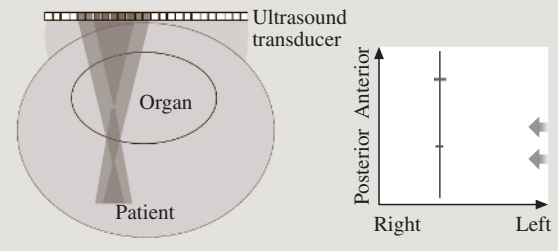


Fig. 21.42 Adjustable focus with large aperture. Using a high = density linear array to generate a large aperture allows electronic adjustment of the focal depth by adjusting the electronic delay of the transmit and receive signals so that the signal at the center transducers is delayed more than the signal at the edge transducers in both transmit and receive. Dual transmit focal depths, indicated by the pair of *arrows* on the right of the image, require that separate transmit pulses be sent into tissue for each of the two focal depths. On receive, the focus can be changed between the time that shallow echoes are received and the time that deep echoes are received

Because a linear array can steer the ultrasound beam, it is possible to form an image of each voxel in the tissue plane from different views, combining the results into a compound image (Fig. 21.43). The algorithm for combining the data from different angles is important in creating a pleasing image.

Phased Array Transducers. By applying a selected time delay to each transducer in an array, the ultrasound beam can be focused or tilted as desired (Fig. 21.39). Phased array transducers (Fig. 21.44) form the ultrasound image by tilting the ultrasound beam using selected delays on each transducer on transmit and receive to achieve the desired tilt. A scanhead with a transducer array of width

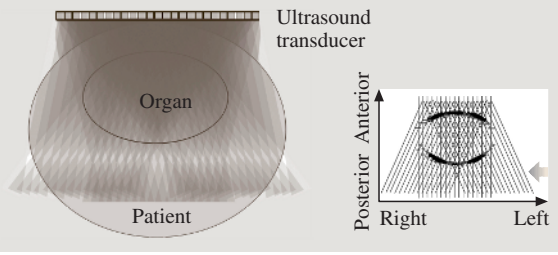


Fig. 21.43 Compound real-time imaging. By tilting the beam patterns to cross during acquisition of the image plane, system can take advantage of the specular backscatter of planar reflectors in the tissue

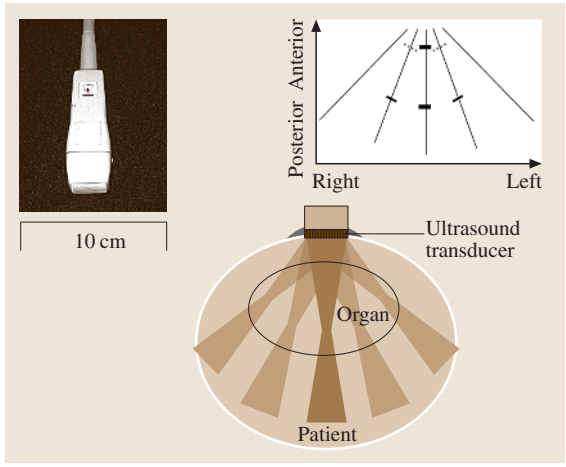


Fig. 21.44 Phased-array scan head. A phased-array scan head is able to use the same aperture for all ultrasound scan line beam patterns, directing the beam along different lines by adjusting the relative phase of the signals to the transducers

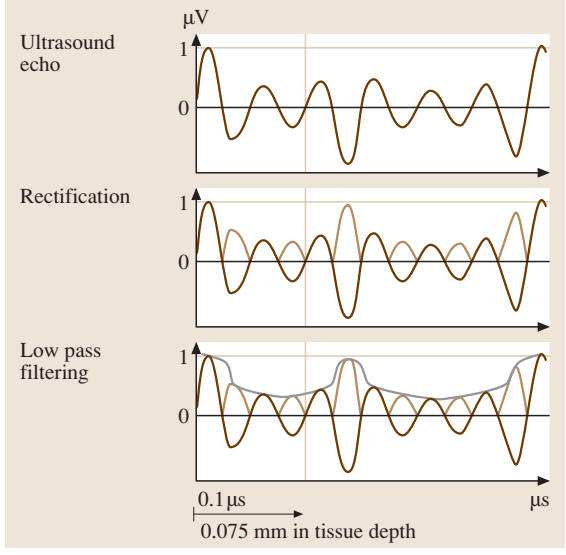


Fig. 21.45 Amplitude demodulation. To create a B-mode image of tissue anatomy, the strength or amplitude of the received ultrasound echo is determined as a function of depth. The RF echo is rectified (converting 5 MHz to 10 MHz) and low-pass-filtered with a 1 MHz cutoff to yield a depth resolution of 0.75 mm. That is effective if the ultrasound transmit burst is shorter than 1 μs. Amplitude demodulation requires one pulse-echo cycle along each ultrasound beam pattern line

Amazon EX1014 Page 874

D can tilt an ultrasound beam to an angle θ in tissue with an ultrasound velocity of c by applying a time delay gradient along the transducer of

$$\frac{tc}{D} = \sin\theta \,. \tag{21.11}$$

The time t is often represented as a function of the phase of the ultrasound ϕ for the ultrasound frequency f_{us}

$$t = \frac{\frac{\phi}{2\pi}}{f_{us}} \,. \tag{21.12}$$

To allow the ultrasound beam to tilt to θ degrees, the width of each transducer element must be less than $\lambda/2$ but the pitch (w = center to center spacing of the elements) must be $\lambda/4$ in order that, at the maximum deflection angle, a single beam pattern is defined when ultrasound intersects the array at angle θ

$$w < \frac{\lambda}{4\sin\theta} = \frac{c}{4 f_{us}\sin\theta} \,. \tag{21.13}$$

A 3 MHz scan head tilting the ultrasound beam 30° to the left and to the right requires a transducer element pitch of 0.3 mm.

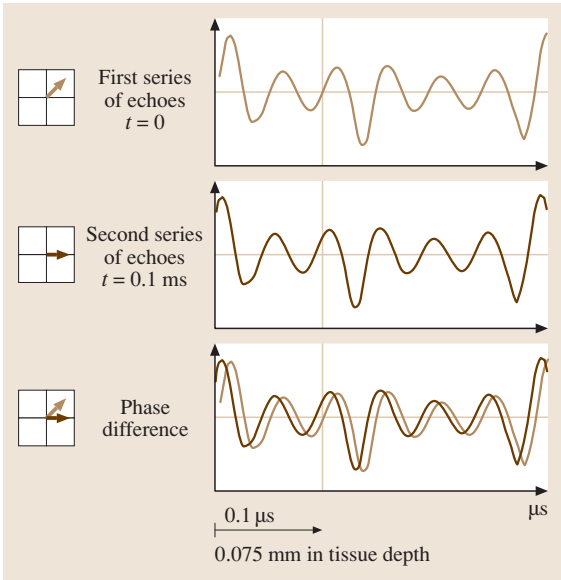


Fig. 21.46 Phase demodulation. To measure blood velocity, or tissue displacement or strain, the change in phase (or echo displacement) between successive pulse-echo cycles at a particular depth must be determined. In this case, a phase change of $\pi/4$ of 5 MHz ultrasound represents an echo time shift of 0.05 μs/8 = 0.006 25 μs, which is equivalent to a displacement of the scatterers toward the transducer of 0.006 25 μs × 0.75 mm/μs = 0.0047 mm in the time between pulse-echo cycles or 0.1 ms, yielding a velocity of 0.047 mm/ms or 0.047 m/s. The Doppler equation, time-domain velocimetry and strain imaging are all based on this computation. In Doppler demodulation, the phase and amplitude of the echo at any depth can be represented as a vector on a phase diagram. These vectors can be represented as complex numbers if the vertical direction is represented as the "imaginary" ($\sqrt{-1}$) direction and the horizontal direction represents the "real" direction

Pulsed Doppler [21.57]

The series of echoes from a single ultrasound transmit pulse provides information about the echogenicity of the tissue along the ultrasound beam pattern at various depths. The echogenicity of the tissue at each depth is determined by amplitude demodulation (Fig. 21.45). The measurement of the velocity of blood or tissue requires pairs of echoes from two (or more) ultrasound transmit pulses along the same beam pattern line, separated by a short time (0.1 ms). By comparing a second echo series from the same transducer beam pattern from a later time to the echo series from the first pulse, motion of the echogenic tissue along the beam pattern can be detected and measured. If the second echo series is identical to the first echo series, then none of the tissue along the ultrasound beam pattern has moved or changed during the period. However, if the second echo series is different, then the tissue has moved or changed. Motion in the direction of ultrasound propagation causes a simple time shift in the echo series, lateral motion or a change in tissue properties results in a loss of coherence between the echoes.

If the tissue has moved toward the ultrasound transducer, then the second echo pattern will arrive earlier after the transmit pulse compared to the first (Fig. 21.46). If the difference is detected by Doppler processing, then the echoes from each region of tissue are characterized by a vector in a complex plane (Fig. 21.46). One rotation of the vector is equivalent to the reflector moving toward the transducer to shorten the round-trip path by one wavelength of the ultrasound. If the ultrasound is 5 MHz, then the wavelength is 320 μm so one rotation represents a shortening of the round trip ultrasound travel distance of 320 μm, or 160 μm each way. A phase rotation of one quarter of a cycle ($\pi/2$ rad) represents 40 μm of motion. Doppler processing cannot tell the difference between 20 and 180 μm of motion because the vector is pointed in the same direction after the 160 μm displacement. This is called aliasing and is responsible for a waveform with the highest forward velocities incorrectly shown as reverse velocities (Fig. 21.47). Doppler

Amazon EX1014 Page 875

systems cannot tell the difference between displacements of 100 μm toward the transducer and 60 μm away from the transducer, because both are represented by the same phase shift. A motion of 20 μm will result in a rotation of the phase vector by 45° ($\pi/4$ rad). Often the interval between the first echo and the second echo is 0.1 ms. The velocity of the tissue in this case is $20/0.1\,\mu m/ms = 200\,mm/s = 20\,cm/s = 0.2\,m/s$.

The Doppler frequency is the phase change of the echo in cycles during the pulse interval divided by the pulse interval. In this case the phase change is 1/8 cycle, so the Doppler frequency is $0.125/0.1\,ms = 1250\,cy/s = 1.25\,kHz$.

Usually, pulsed Doppler systems are used to measure blood velocity. Unfortunately, it is rarely convenient to align the ultrasound beam pattern along the axis of a blood vessel so typically the angle between the ultrasound beam pattern and the vessel axis is 60°. This is called the Doppler angle. The ultrasound system can only measure the projection of the blood velocity vector along the ultrasound beam. That projection is the cosine of the angle between the velocity vector and the ultrasound beam. The Doppler equation relating velocity to the ultrasound echo is:

$$f = \frac{2 v_f f_{us} \cos\theta_D}{c} \quad \text{or} \quad v_f = c\frac{\frac{f}{f_{us}}}{2\cos\theta_D} . \qquad (21.14)$$

where f is the Doppler frequency that you hear, the factor of 2 accounts for the round-trip of the ultrasound, v_f is the velocity of the blood, f_{us} is the ultrasound frequency, θ_D is the angle between the blood velocity and the ultrasound beam, and c is the speed of ultrasound in tissue. The form

$$\frac{v_f}{c} = \frac{\frac{f}{f_{us}}}{2\cos\theta_D} \qquad (21.15)$$

emphasizes the relationship between the ratio of the blood velocity to the ultrasound velocity and the Doppler frequency to the ultrasound frequency. Since the ultrasound wavelength $\lambda = c/f_{us}$,

$$v_f = \frac{f\lambda}{2\cos\theta_D} . \qquad (21.16)$$

To provide a measurement of blood velocity, ultrasound instruments have a cursor on the screen to align with the axis of the vessel under examination (Fig. 21.48). The associated Doppler data, here represented as a spectral waveform, is provided with a scale showing the angle-adjusted velocity based on the projection of the measured velocity onto the vessel axis. Although this adjustment is theoretically justified by the geometry, unfortunately, it does not provide a useful velocity value. If a Doppler angle of zero degrees is used, then the value is correct, however, if other angles between the ultrasound beam and the presumed velocity vector are used, then the value of the velocity is not correct. The reason is that blood velocity vectors are rarely aligned with the axis of the artery. Instead, blood velocities form helical patterns along the artery [21.58–60]. Angle-adjusted Doppler velocity measurements are systematically higher as the Doppler angle increases from zero to 70 degrees or more. Although at a Doppler angle of zero, velocities are accurately used to determine both the volume flow rate and the Bernoulli pressure drop, neither result is correct when Doppler angles other than zero are used [21.61–63].

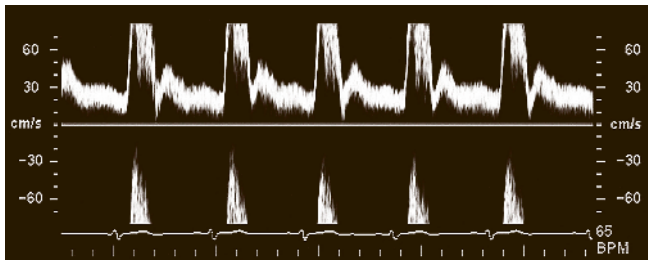


Fig. 21.47 Aliasing. Blood velocities that cause high phase shifts (greater than 180° or π radians) between pairs of pulsed Doppler echoes cause the Doppler system to mistake high-velocity forward flow for reverse flow. Continuity in blood flow allows the examiner to detect this error

Pulsed Doppler Range of Measurement. The highest Doppler frequency shift that can be measured with pulsed Doppler occurs when the phase change of the echoes during the interval between echo acquisitions [pulse repetition interval (PRI)] is 180° (half a cycle). The Doppler frequency at that condition is half of the pulse repetition frequency (PRF, the Doppler sampling frequency). This is called the Nyquist limit, after Harry Nyquist. This establishes the highest velocity that the Doppler can detect, although since the signal is analyzed as a complex number with twice the data (half as a "real" number and half as an "imaginary" number) Doppler frequencies between $+0.5$ PRF and -0.5 PRF are reported. As long as there is no reverse flow the range of Doppler frequencies available is equal to the PRF. If the phase vector is characterized by a 2 bit complex number into four values, then the smallest phase change that can be classified is 1/4 cycle. So, a 5 MHz Doppler system with the data recorded as 2 bits, and a PRF of 10 kHz, can only report four Doppler frequency ranges,

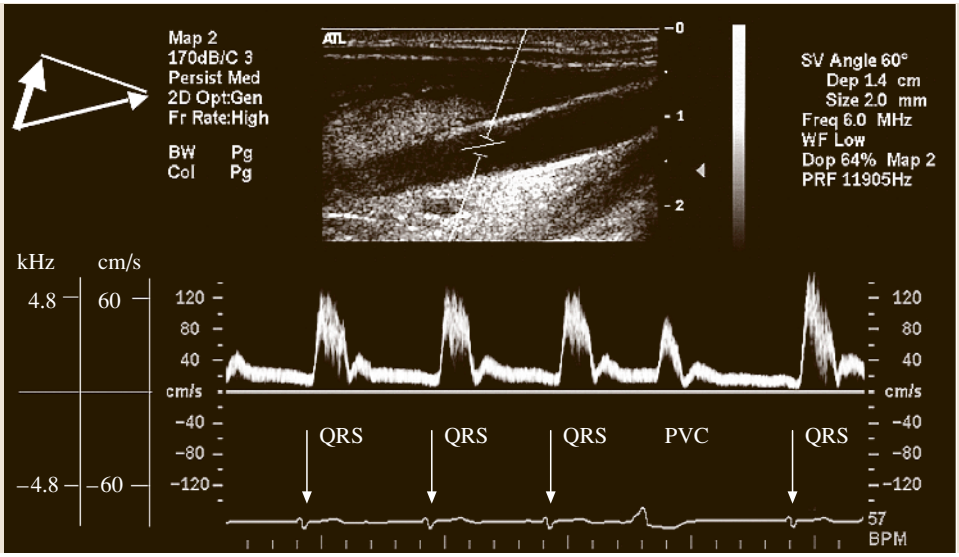


Fig. 21.48 Duplex Doppler image and waveform of a carotid artery. The axis of the ultrasound beam pattern used for Doppler is shown crossing the artery at an angle of 60° to the vessel axis. A cursor is aligned parallel to the vessel axis to provide θ_D for the Doppler equation. The Doppler frequency, which is produced by the speakers of the instrument, is shown by a scale drawn on the *left*. The scale showing the speed at which the blood is approaching the transducer is drawn adjacent to the frequency scale. The instrument computes the scale on the screen by dividing the cosine of the Doppler angle into the approach speed. Along the lower edge of the image is the electrocardiograph tracing (ECG) showing the "QRS complex" (small up and down) indicating the onset of ventricular contraction in each cardiac cycle. The lump in the trace 200 ms later is the T wave indication the end of ventricular contraction. The fourth beat is too early and the QRS has an odd shape. This is a premature ventricular contraction (PVC) originating from an abnormal spot on the ventricle. The PVC causes the ventricle to contract before filling is complete, producing a low-volume contraction, and low systolic blood velocities. The ventricle overfills before the next normal contraction, causing elevated systolic velocities

-1.25 to $+1.25$ kHz, $+1.25$ to $+3.75$ kHz, $+3.75$ to $+6.25$ kHz, and $+6.25$ to $+8.75$ kHz. -1.25 kHz is equal to $+8.75$ kHz. Such a Doppler system would probably classify -1.25 to $+1.25$ kHz as no motion. By increasing the number of bits in the digitizers representing the phase vector, the resolution of the phase shift is increased, and the number of possible measured velocities is increased. This does not increase the Nyquist limit.

Another way to increase the number of velocities that can be measured is to increase the number of pulse-echo cycles or Doppler phase samples that are gathered. The number of samples used to resolve the velocity is called the ensemble length or packet length. In a typical spectral waveform pulsed Doppler system, Doppler samples will be gathered from a voxel at a PRF of 12.8 kHz for 10 ms to provide 128 Doppler samples for frequency analysis. The Nyquist limit (upper frequency limit) is 6.4 kHz. The lowest frequency that can be detected is 100 Hz, one full Doppler cycle in the 10 ms acquisition period. If the signals are digitized into 2 bit complex numbers, then each of the 128 possible frequencies can be tested to see whether each is present or absent, but no information is available about the strength of each frequency signal. To determine the strength of each frequency, the signals must be digitized into numbers with more binary bits. Spectral waveform analysis assumes that there are many Doppler frequencies present in the Doppler signal from the voxel of interest.

Pulsed Doppler Pulse-Pair analysis. The Doppler frequency shift can be measured from just two echo samples separated by a known time. The assumption is that only a single Doppler frequency is present in the signal. The phase change divided by the sample interval time provides the Doppler frequency. The upper limit of the Doppler frequency is the Nyquist limit. To resolve the Doppler frequency, the phase angle change must be resolved. Representing the phase angle by a two bit complex number only resolves the phase angle into four divisions. For 5 MHz ultrasound, displacements of

Amazon EX1014 Page 877

40 μm can be resolved. For a 10 kHz PRF sample frequency, this allows a velocity of ±40 cm/s (±0.4 m/s) to be detected. However, if the phases are represented by 10 bit complex numbers, the phase angles can be resolved into 1000 divisions and the displacement and velocities can also be resolved. With 10 bit quadrature digitizers, velocities between plus and minus 80 cm/s can be measured with a resolution of 0.8 cm/s, equivalent to displacements of 80 nm during the pulse interval.

Pulsed Doppler Wall/Clutter Filtering. The measurement of blood velocity with ultrasound is complicated by the containment in vascular walls. The echogenicity of the blood is 60 dB less than the echogenicity of the walls at 5 MHz. The blood cells are Rayleigh scatterers because they are smaller (7 μm) than the wavelength of ultrasound (300 μm for 5 MHz). The strength of the echo from blood increases with the fourth power of the frequency. However, frequency-dependent attenuation of the ultrasound by overlying tissues limits most Doppler ultrasound studies to frequencies of 2–6 MHz. 5 MHz produces the strongest echoes for vessels under 2 cm of tissue, deeper vessels, and those behind the skull require 2 MHz ultrasound waves to penetrate these tissues [21.64–66].

In many examinations, Doppler measurements from blood adjacent to the wall are desired, so Doppler echoes from the voxel next to the wall will include strong wall echoes as well as weak echoes from the adjacent blood. Unfortunately, vascular walls move with velocities near 2 mm/10 ms or 0.2 m/s (20 cm/s). The strong echoes from these moving walls must be rejected to reveal the adjacent blood velocity, which may be only slightly higher than 20 cm/s. The strong echoes from the wall are called clutter. 10 bits of the system dynamic range are required to handle the clutter and the blood echoes must be resolved above that, so most ultrasound systems now utilize a 14–16 bit dynamic range, which is achieved by combining the outputs of separate digitizers (one for each transducer) and digitizing faster than the required four times the ultrasound center frequency, then filtering the result for processing.

Pulsed Doppler Spectral Analysis. To measure the Doppler frequency shift, any method of frequency analysis can be used. Zero-crossing rate meters and zero-crossing time interval histograms have been replaced with real-time spectrum analyzers. Although the time-compression analyzer produces identical results, the popular fast Fourier transform (FFT) analyzer is used to produce spectral waveforms of Doppler signals

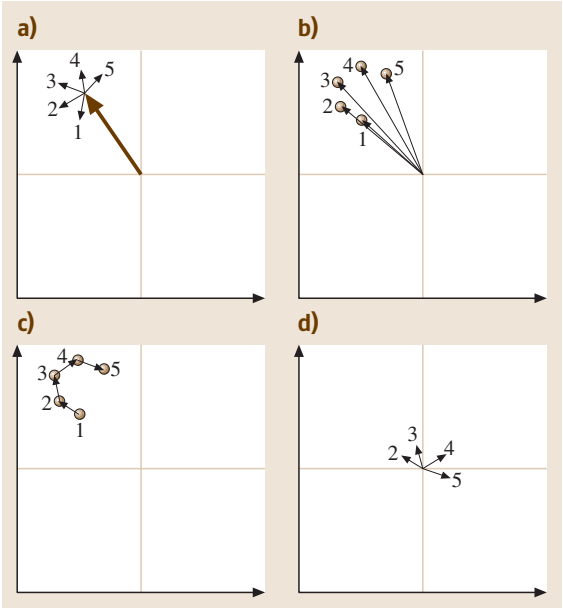


Fig. 21.49a–d Autocorrelation clutter (high-pass) filter. By representing each pulse-echo from a particular depth as a point on a phase diagram, the trajectory of the echo phase in time can be plotted. **(a)** The Doppler echo comes from two types of reflectors, strong reflecting stationary structures (*clutter*) and weak reflecting moving blood cells. In each of the five pulse-echo cycles (ensemble or packet of 5), the contribution from the strong reflectors is the same. The moving blood cells reflect the same amplitude (vector length) but a different phase during each pulse-echo cycle. **(b)** The ultrasound receiver cannot differentiate the clutter contributed by the stationary structures from the signal contributed by the moving blood cells. High-pass filter. **(c)** By taking the difference between successive results, the effect of the clutter is eliminated. **(d)** By centering the resultant vectors on a new phase diagram, the phase angle change between each pulse-echo cycle can be used to compute the blood velocity

(Figs. 21.47, 21.48). The FFT analyzer is coupled to the pulsed Doppler so that the FFT analyzer receives a quadrature pair of data from the depth of interest for each pulse-echo cycle. The advantage of the spectral analysis is that if the voxel contains several flow velocities, the FFT spectrum will display the Doppler frequency for each velocity. Unfortunately, in blood flow, eddy oscillation frequencies up to 500 Hz occur; the typical FFT length of 10 ms (100 spectra per second) is only capable of resolving eddies of 50 Hz or lower. When a voxel contains higher-frequency eddies, the re-

sult appears as spectral broadening or as harmonics on the spectral waveform.

Pulsed Doppler Autocorrelation. Autocorrelation is a frequency analysis method which is an extension of pulse-pair analysis. Autocorrelation analysis assumes that there is only one valid Doppler frequency in the blood flow. Autocorrelation analysis identifies that frequency. However, to reject the stationary and slow moving strong echoes of the wall, a clutter filter must be applied. Clutter is a stationary echo superimposed on the changing Doppler echo. The analysis is based on an ensemble of N echoes, usually eight echoes. The difference between adjacent pairs of echoes n and $n-1$ (for $n=2,\ N$) creates a new group of $N-1$ vectors (Fig. 21.49). By taking the average of the change in angles, a Doppler velocity can be obtained. Sometimes two filtering steps are required rather than one. Such filters are called finite impulse response (FIR) and infinite impulse response (IIR) filters.

Time–Domain Velocimetry. The Doppler equation contains the Doppler frequency and the phase angle. Doppler signals are subject to aliasing when the between-pulse phase angle change ($\Delta\phi$, a circular variable) has a value greater than π. In addition, echoes from deep tissues have a lower center frequency than echoes from shallow tissues, causing confusion about the correct value of the ultrasound frequency f_{us} to use in the equation.

$$\frac{v_f}{c} = \frac{\frac{\Delta\phi}{\Delta t}}{f_{us}\, 2\cos\theta_D}\,. \qquad (21.17)$$

Of course, the center frequency of deep echoes is still within the bandwidth of the transmitted ultrasound; the center frequency is depressed because of the greater attenuation of the higher-frequency portions of the transmitted bandwidth.

Since bandwidth is inversely proportional to pulse duration and pulse duration is related to spatial resolution, it is desirable to have a short pulse duration so that spatial resolution is as small as possible. This means broadband pulses, permitting depression of the center frequency of the echo. The desire for improved resolution and the desire to avoid aliasing combine in the use of time-domain velocimetry as an alternative to Doppler.

$$\frac{v_f}{c} = \frac{\frac{\Delta\tau}{\Delta t}}{2\cos\theta_D}\,. \qquad (21.18)$$

In time-domain velocimetry, two successive ultrasound echoes are compared to see if a time shift ($\Delta\tau$) has occurred in the echo in the intervening time (Δt), which indicates tissue motion. By measuring the time shift, the velocity can be calculated. Of course, clutter from solid tissue in the voxel may provide a substantial stationary echo superimposed on the small shifted echo. To remove the clutter echo, subtraction of adjacent pairs

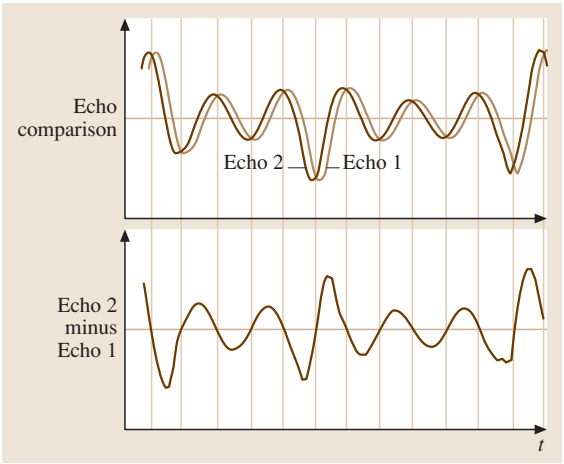


Fig. 21.50 Time-domain demodulation removing clutter. To remove clutter in time-domain demodulation, the RF echo signal from each pulse-echo cycle is subtracted from the preceding RF echo signal to yield the residual due to tissue motion

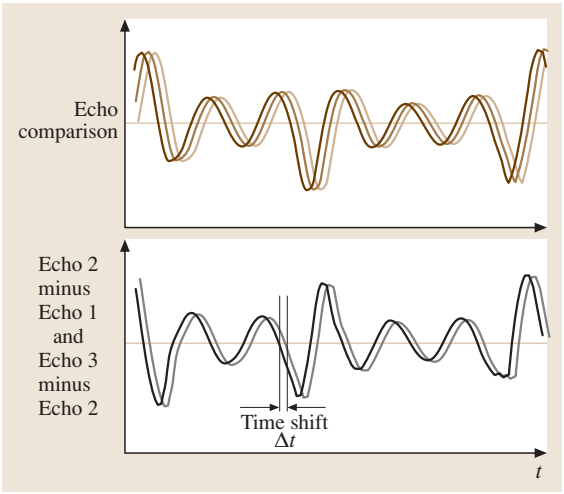


Fig. 21.51 Time-domain demodulation measuring the velocity component. By measuring the time shift between successive echo pairs, multiplying by half the ultrasound velocity and dividing by the pulse repetition interval time, the velocity of the tissue along the ultrasound beam results

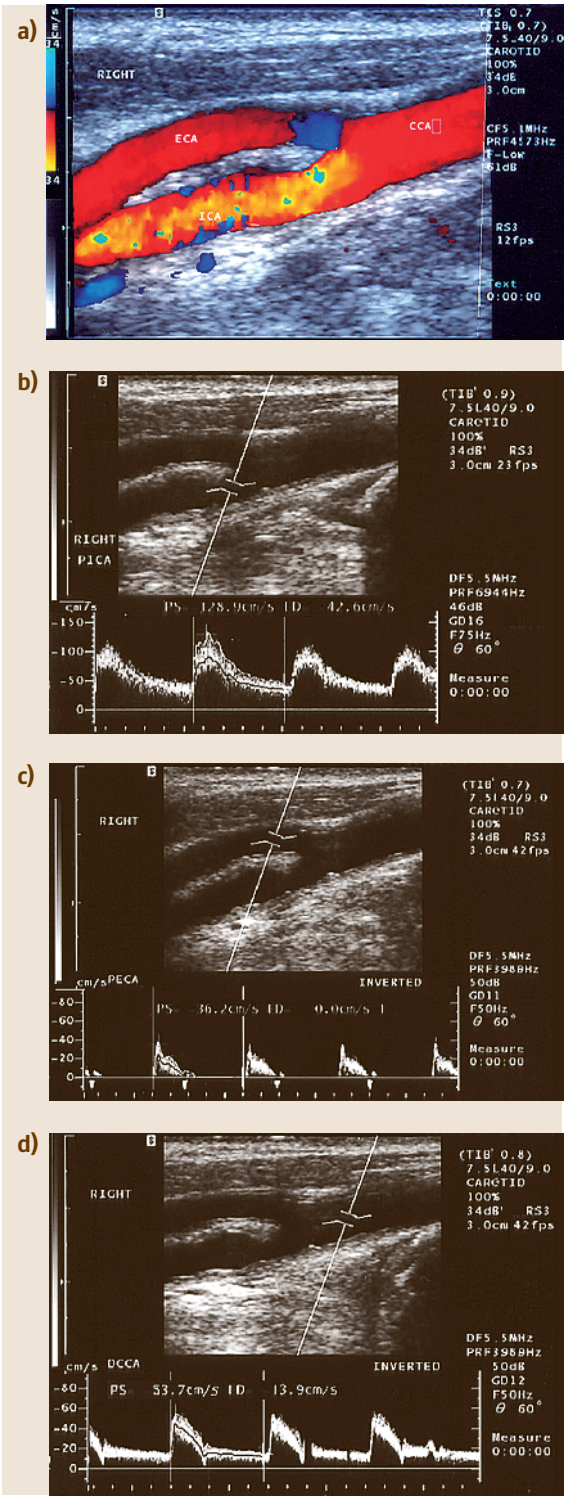


Fig. 21.52a–d Ultrasound images from a conventional carotid artery Doppler examination in the neck [21.67]. *Orientation*: a linear array transducer located at the top of each image acquires the 128 vertical scan lines in sequence across the 35 mm-wide image. Two of the monochrome B-mode images **(c)** and **(d)** were acquired in 24 ms (42 frames per second) so the scan lines progress across the image at a speed of (35/24) 1.46 m/s, a speed similar to blood velocity. The color flow image **(a)** was acquired at 12 fps; the other monochrome image **(b)** was acquired at 23 fps. *Anatomy*: **(a)** ECA external carotid artery supplying the face, ICA internal carotid artery supplying the brain, CCA common carotid artery from the heart via the Brachiocephalic trunk. **(a)** Color Doppler image of the carotid bifurcation. The color scale on the left of the image shows that forward flow is shown as *red* and reverse flow is shown as *blue*. The flow directions are with respect to the transducer array at the top of the image rather than with respect to the heart to the right of the image. The blue region at the origin of the external carotid artery shows that the vertical component of the velocity vector is pointing toward the ultrasound transducer rather than showing that flow is reversed toward the heart. Notice that this region is bounded by black bands indicating *zero* velocity toward the transducer. The *blue spots* in the lumen of the internal carotid artery are aliasing (Fig. 21.47). Notice that each of those regions is surrounded by an *orange* color indicating high blood velocity components with respect to the transducer. *Doppler waveform orientation*: each of the three Doppler waveforms is acquired from an ultrasound beam pattern that is tilted at an angle with respect to the transducer, selected by the examiner to form an angle of 60° with respect to the artery axis. **(b)** *lower* Normal Doppler spectral waveform from the internal carotid artery with high diastolic flow. Because the brain continuously requires 10% (10 ml/s) of the cardiac output (100 ml/s), the peripheral resistance is low so blood flows to the brain during systole and diastole. **(c)** *lower* Normal Doppler spectral waveform from the external carotid artery with low diastolic flow. Because the face at rest requires little blood flow but when active (chewing, blushing) requires more blood flow, at rest the arteriolar sphincters are constricted, only accepting blood flow during systole. When the arteriolar resistance decreases, the external carotid also has flow during diastole. **(d)** *lower* Normal Doppler spectral waveform from the common carotid artery which has the combined internal and external flow waveform. A pulsatility index showing the ratio between the diastolic and systolic velocities is sometimes used to classify waveforms

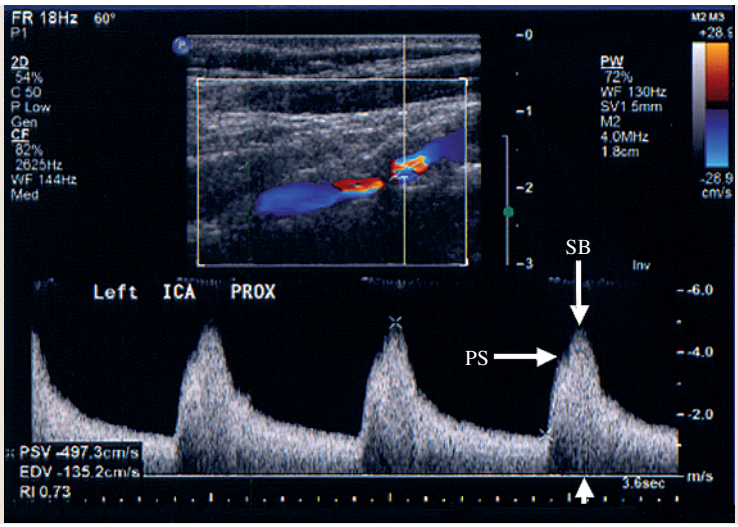


Fig. 21.53 Stenotic internal carotid artery. Even after 25 years of exploring the meaning of carotid artery spectral waveforms, there is no agreement on how the data should be interpreted. Although this color Doppler image shows an interruption in the color flow image of the lumen, the interruption is not due to an arterial occlusion, but to a local region of high ultrasound attenuation which causes the color Doppler signal to drop out. Thus, the diameter of the color Doppler streak cannot be used to assess the lumen. In this case, the vessel walls cannot be easily seen. A stenosis is present at the Doppler sample volume as indicated by the high velocity. The spectral waveform is confusing. The highest Doppler frequency shift which was measured by the examiner 220 ms after the onset of systole at 497.3 cm/s is probably not accurate because the measurement is at the time of greatest spectral broadening (SB), indicating eddies in the flow. Use of the Doppler equation is not appropriate for these eddy velocities because the heading of the vectors is not known. The true peak systolic (PS) velocity probably occurs 100 ms after the onset of systole where the measurement is near 4 m/s and the Doppler equation is probably accurate if the velocity vector is assumed to be parallel to the artery axis. An appreciating of the difference between the believed accuracy of 497.3 cm/s versus 'about 4 m/s' is important because of the clinical decisions that result. The PSV is used to decide whether a patient with symptoms of mini-stroke should have surgery. In this case, the end diastolic velocity (EDV) is measured at 135.2 cm/s. An EDV value exceeding 140 cm/s in a single measurement would cause a vascular surgeon to schedule the patient for carotid endarterectomy surgery even if the patient has no symptoms

of echoes (Fig. 21.50) provides the difference waveform containing the motion information. Then the time shift between difference waveforms can be measured (Fig. 21.51).

A comparison between the time-domain equation and the Doppler equation shows that $\Delta\tau = \Delta\phi / f_{us}$. This indicates that it is the demodulator frequency, not the frequency of the echo, that is relevant for the Doppler equation; the former must be close (and is usually equal) to the transmitted center frequency, i. e., the frequency required in the Doppler equation. There are substantial differences between the Doppler and time-domain systems. First the time-domain system usually transmits a broadband, short ultrasound pulse. The signal-to-noise ratio of the echo is determined by the transmit pulse energy. To achieve the same energy in a short pulse, the temporal peak power must be high, leading to a high acoustic pressure. When analyzing the time shift between the echoes, it is easiest to test for the time-lagged correlation by stepping one digitizing interval at a time. So, if the echo is digitized at a rate of 20 MHz, the digitizing interval is 50 ns ($\Delta\tau$), equivalent to a tissue displacement of 0.0375 mm. If the pulse repetition frequency is 10 kHz, so the inter-pulse interval is 100 µs (Δt), the velocity resolution is 38 cm/s (0.4 m/s). Time-domain systems usually digitize at a higher frequency and usually interpolate the wave between the measurements to improve velocity resolution. For large time shifts, when noise is present, the time domain system usually has trouble with peak hopping, associating the wrong cycles in the wave when determining the time shift. This results in a measurement error identical to

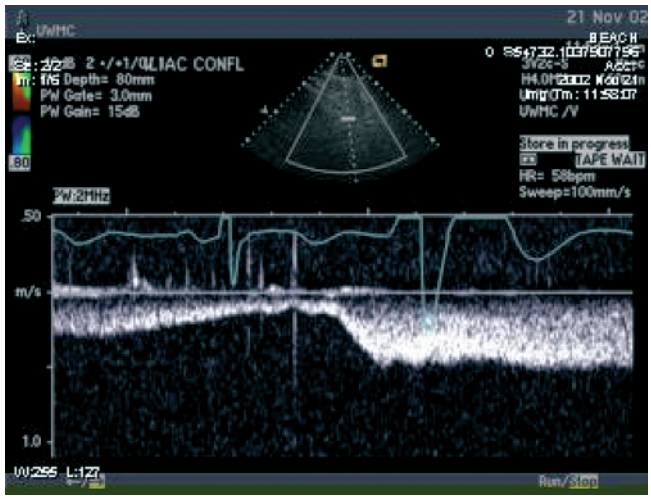


Fig. 21.54 Doppler spectral waveform signals from the inferior vena cava showing augmented flow. Doppler spectral waveforms show velocity (*vertical axis*) versus time (*horizontal axis*). In addition to the difference in flow direction, venous flow is differentiated from cardiac flow because the venous flow varies with respiration and with manual compressions of the tissues. Here the sudden increase in venous flow in the distal inferior vena cava was created by releasing an occlusion of the venous outflow from the legs. ECG tracing (*green*) shows cardiac timing. The heart beat to the right of center is due to a premature ventricular contraction, an abnormal heart rhythm which does not affect the venous flow

aliasing. So, the advantages of time-domain processing over Doppler processing, while promising, have not been fulfilled.

Clinical Applications of Ultrasound Velocimetry

The most common application of pulse Doppler ultrasound spectral waveform analysis is the examination of the carotid arteries [21.68] for stenoses that might lead to stroke (Figs. 21.52, 21.53). The measurement of high systolic (> 125 cm/s) carotid blood velocity is associated with a 20% chance of stroke in two years in a patient reporting stroke-like symptoms

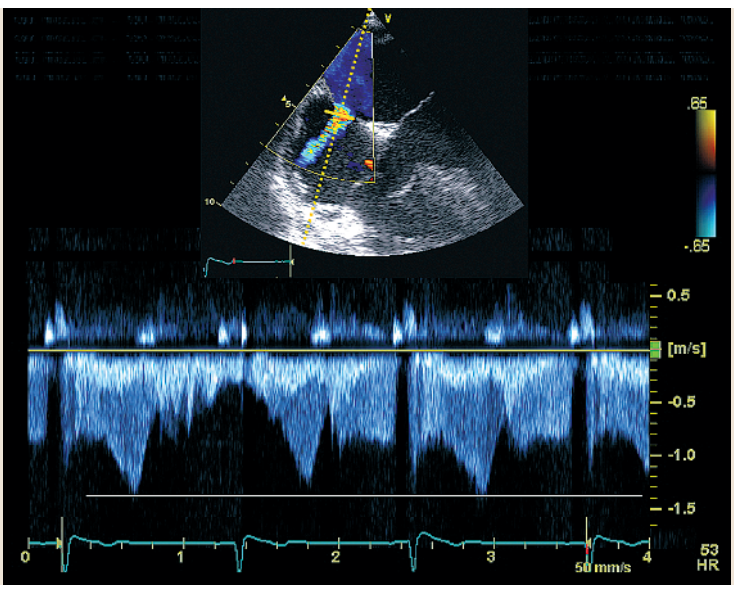


Fig. 21.55 Duplex echocardiogram of a heart with an atrial septal defect. The jet of blood from the left atrium to the right atrium can easily be seen on the color flow image. The Doppler spectral waveform taken at the center of the jet orifice at a Doppler angle of 25° to the jet shows velocities as high as 1.4 m/s when the inter-atrial pressure difference is the greatest. Correcting the velocity for the Doppler angle ($\cos 25 = 0.906$) increases the velocity value to 1.54 m/s. Cardiologists usually attempt to orient the Doppler transducer to align the jet with the Doppler ultrasound beam to achieve a *zero* Doppler angle. The pressure difference between the chambers can be computed using the Bernoulli equation for blood velocity $\Delta p[\text{mmHg}] = 4\,[\text{mmHg}/(\text{m/s})^2]\,\text{V}^2$. The pressure difference here is near 9 mmHg (courtesy Frances DeRook and Keith Comess)

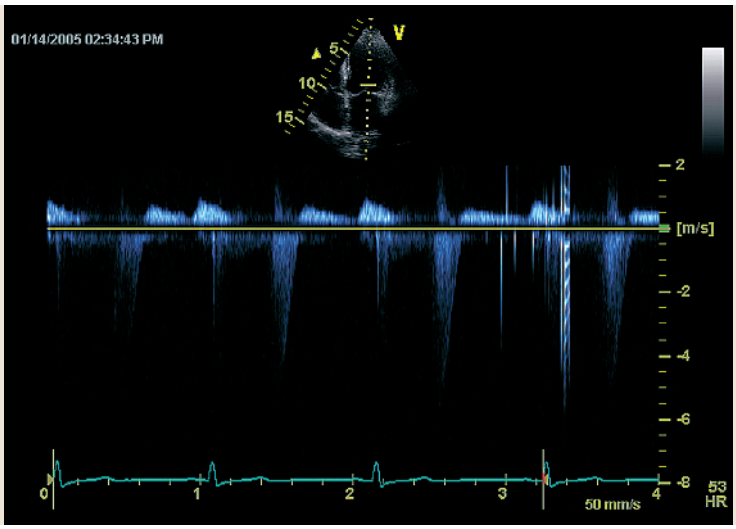


Fig. 21.56 Vibration from the mitral valve. Signals from blood flow are mixed with signals from solid tissue motion. Here, the signal of solid tissue motion is suppressed by the wall filter, which suppresses clutter at frequencies below 400 Hz. However, in this case with the Doppler sample volume in the mitral orifice, as the mitral valve closes for the fourth time, the valve vibrates at a frequency of 2 kHz for a period of 8 ms causing the harmonic (Bessel function) spectral pattern with seven harmonic terms, indicating that the amplitude of the vibration is high (courtesy Frances DeRook and Keith Comess)

(TIA = transient ischemic attack, RIND = reversible ischemic neurological deficit). The measurement of a high-end diastolic velocity (> 140 cm/s) in any patient is associated with a 20% chance of stroke in two years from emboli released from the carotid stenosis. Such findings, by themselves, can direct a patient to treatment that will prevent devastating disability.

Examination of the veins of the legs with Doppler for the detection of deep venous thrombosis (DVT) is another popular examination that can lead to life sav-

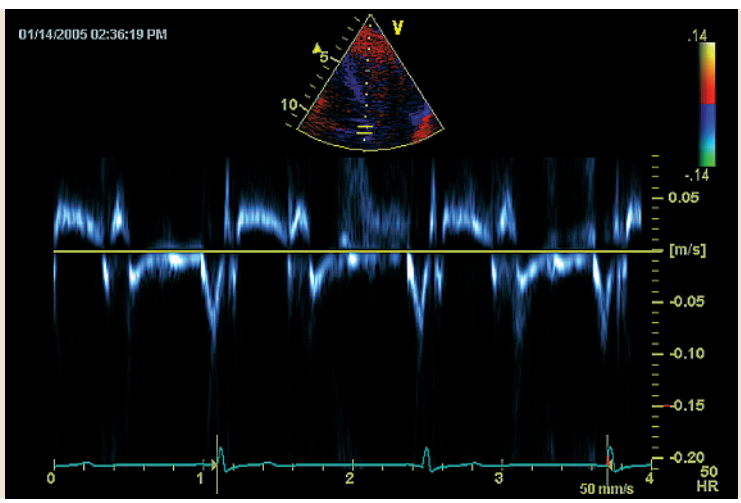


Fig. 21.57 Duplex echocardiogram tissue Doppler of myocardium. Like velocity measurements from blood flow, velocities of solid tissues can also be measured. Low velocities of the intra-ventricular septum of the heart indicates abnormal muscle contraction (courtesy Frances DeRook and Keith Comess)

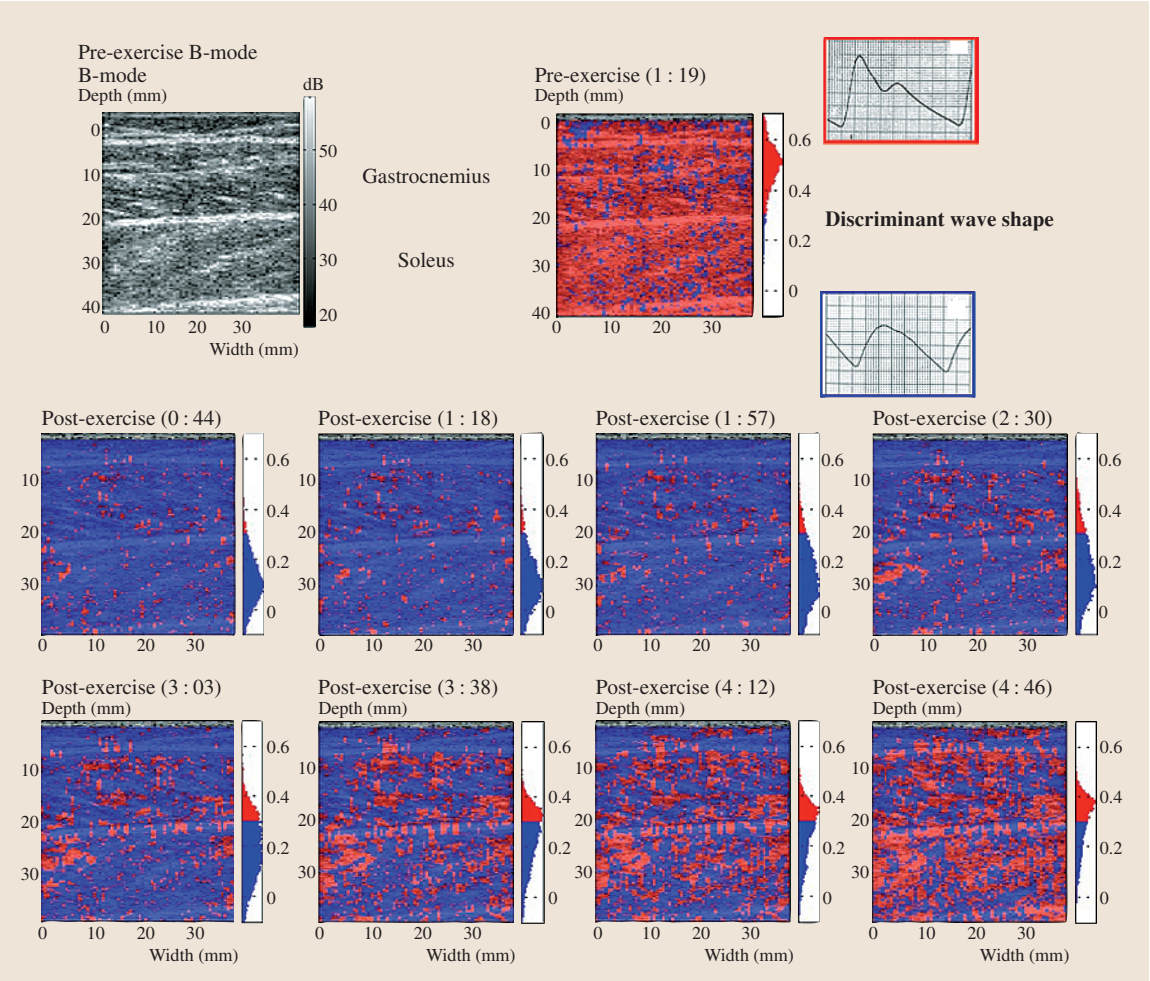


Fig. 21.58 Tissue strain in calf muscles. During tissue perfusion, different strain waveforms indicate resting perfusion (red) with high peripheral (arteriolar) resistance or hyperemic perfusion (blue) with low peripheral resistance. In this case, the calf muscle was imaged at rest, and then at intervals for a period of 5 min following a toe stand exercise. Immediately after exercise, the peripheral resistance is low to make up for the metabolic oxygen deficit after exercise. Five minutes after exercise, most of the muscle tissue has returned to the resting state (Courtesy John Kucewicz)

ing therapy. Because venous pressures are near 5 mmHg in the supine patient, normal venous flow is associated with respiration and minor local compression rather than pulsatile with the cardiac cycle (Fig. 21.54). Pulmonary embolus from venous emboli released from a DVT may account for one-third of sudden cardiac death cases [21.69]. A venous Doppler examination failing to show spontaneous venous flow with respiration indicates a venous obstruction that may be a DVT that could progress to threaten life if not treated with anticoagulation.

Cardiac Doppler examinations for blood flow are perfectly suited to characterize narrow holes between the cardiovascular structures in the chest (Fig. 21.55). By measuring the velocity of blood flow jets and applying the Bernoulli equation, the instantaneous pressure difference between the associated chambers of the heart can be measured. This works for all heart valves and wall defects so the method can be used to identify pulmonary artery hypertension as well as cardiac valve failure.

Although blood velocity through valves can be measured (Fig. 21.56), the dynamics of solid structures such