An international comparison of long-term average speech spectra

Denis Byrne, Harvey Dillon, and Khanh Tran
National Acoustic Laboratories, 126 Greville St., Chatswood, 2067 NSW, Australia

Stig Arlinger
University Hospital, Linkoping, Sweden

Keith Wilbraham
University of Manchester, Manchester, United Kingdom

Robyn Cox
Veterans Affairs Medical Center, Memphis, and Memphis State University, Memphis, Tennessee

Bjorn Hagerman
Karolinska Institute, Stockholm, Sweden

Raymond Hetu
GAUM, Montreal, Canada

Joseph Kei and C. Lui
Special Education, Kowloon, Hong Kong

Jurgen Kiessling
University Hospital, Giessen, Germany

M. Nasser Kotby, Nasser H. A. Nasser, and Wafaa A. H. El Kholy
Ain Shams University, Cairo, Egypt

Yasuko Nakanishi[a)]
Gakugei University, Tokyo, Japan

Herbert Oyer
Ohio State University, Columbus, Ohio

Richard Powell
Taranaki Hospital, New Plymouth, New Zealand

Dafydd Stephens, Rhys Meredith, and Tony Sirimanna
Welsh Hearing Institute, Cardiff, Wales

George Tavartkiladze and Gregory I. Frolenkov
Research Center for Audiology, Moscow, Russia

Soren Westerman and Carl Ludvigsen
Widex, Vaerloese, Denmark

(Received 4 November 1993; accepted for publication 1 July 1994)

The long-term average speech spectrum (LTASS) and some dynamic characteristics of speech were determined for 12 languages: English (several dialects), Swedish, Danish, German, French (Canadian), Japanese, Cantonese, Mandarin, Russian, Welsh, Singhalese, and Vietnamese. The LTASS only was also measured for Arabic. Speech samples (18) were recorded, using standardized equipment and procedures, in 15 localities for (usually) ten male and ten female talkers. All analyses were conducted at the National Acoustic Laboratories, Sydney. The LTASS was similar for all languages although there were many statistically significant differences. Such differences were small and not always consistent for male and female samples of the same language. For one-third octave bands of speech, the maximum short-term rms level was 10 dB above the maximum long-term rms level, consistent across languages and frequency. A "universal" LTASS is suggested as being applicable, across languages, for many purposes including use in hearing aid prescription procedures and in the Articulation Index.

PACS numbers: 43.70.Gr, 43.72.Ar, 43.66.Ts

[a)]Currently at the University of Tsukuba.

Sony v. Jawbone
U.S. Patent No. 8,321,213
Sony Ex. 1009

INTRODUCTION

Representations of the long-term average spectrum of speech (LTASS) have various acoustical and audiological applications. In rehabilitative audiology the LTASS is widely used in the prescription and evaluation of hearing aid fittings. It is used in many hearing aid prescription procedures either in the derivation of the prescriptive formula (e.g., Berger *et al.*, 1977; Seewald *et al.*, 1985; Byrne and Dillon, 1986) or in calculating the prescription for the individual client (e.g., Cox, 1988; Skinner, 1988). The Articulation Index (ANSI, 1969), which has many current or potential uses including the evaluation of hearing aid fittings (e.g., Pavlovic, 1989), depends on using the LTASS. Although there are many published measurements of the LTASS, no particular set of values is universally accepted. Indeed, the different hearing aid selection procedures use various sets of values (Skinner, 1988) and, although the Articulation Index Standard includes an idealized LTASS, this idealized LTASS is not usually recommended in audiological applications (Pavlovic, 1989).

The majority of published measurements of the LTASS are for the English language, as spoken in the U.S.A. (Dunn and White, 1940; Stevens *et al.*, 1947; Rudnose *et al.*, 1958; Benson and Hirsh, 1953; Harris and Waite, 1965; Niemoller *et al.*, 1974; Pearsons *et al.*, 1977; Cox and Moore, 1988; Cornelisse *et al.*, 1991; Stelmachovicz *et al.*, 1993), Australia (Byrne, 1977; Byrne and Dillon, 1986), England (Boothroyd, 1967), or other countries (Tarnoczy, 1971). There are also measurements for other languages which include: German, Hungarian, Italian, Russian (Tarnoczy, 1971), Swedish (Aniansson, 1974; Liejor, 1989), Danish (Dalsgaard and Pedersen, 1966), Finnish (Kiukaanniemi, 1980; Kiukaanniemi *et al.*, 1982), Mandarin (McCullough *et al.*, 1993), French and Dutch (Harmegnies and Landercy, 1985), Polish (Zalewski and Majewski, 1971), and Spanish (Banuls-Terol, 1971). All measurements of the LTASS are similar but small differences occur between English as spoken in different countries and among different languages. However, it is not clear whether these differences are real, because there are large differences among individuals (Dunn and White, 1940; Byrne, 1977; Kiukaanniemi *et al.*, 1982) and many of the studies have used only small subject groups. Furthermore, some studies (Tarnoczy and Fant, 1964; Dalsgaard and Pedersen, 1966; Tarnoczy, 1971; Niemoller *et al.*, 1974; Zalewski and Majewski, 1971; Banuls-Terol, 1971) do not permit any estimate of individual variability because the analyses have been based on a "chorus" of all (or groups of) subjects combined.

By comparing three studies that did use substantial subject groups (at least 20), Cox and Moore (1988) concluded that there probably are small differences in the LTASS of American and Australian speech. From examining samples of Swedish, Hungarian, and German, Tarnoczy and Fant(1964) concluded that there were significant differences among these languages in the midfrequency region (700–1500 Hz for males, 1000–2000 Hz for females). They attributed this difference to the relative occurrence of vowels with second formants in this region. Harmegnies and Landercy compared Dutch and French speech spectra of 20 male talkers who spoke both languages. They concluded that individual talker differences accounted for most of the variability in spectra but that there probably were small language differences which did not exceed 5 dB in any frequency region. They suggested that the differences probably arose from differences in phoneme distributions for the two languages. One particular effect, which occurred around 1000 Hz, was attributed to the existence of nasalized vowels in French contrasting with the absence of any such vowels in Dutch. On the other hand, from a comparison of the LTASS of French, Dutch, English, Italian, and Danish (two male and two female talkers for each language), Pavlovic *et al.* (1991) concluded that there were no significant effects of either sex or language. Overall, it appears that the LTASS may vary with language but the issue is still not resolved.

In addition to subject variation, small differences could occur among different sets of measurements because of variations in measurement techniques. For example, Dunn and White (1940) commented on some earlier measurements that contained an artifact, apparently due to close talking conditions. In the various studies there have been differences in the recording or analysis conditions and these differences may be responsible for some small differences in results. One important procedural variable is the angle of incidence of the recording microphone to the talker's mouth. The majority of studies have used 0° incidence (i.e., microphone directly in front of mouth) recorded in anechoic or approximately equivalent conditions. Some, however, have recorded a "chorus" of talkers with the microphone placed in the diffuse far field (Tarnoczy and Fant, 1964; Dalsgaard and Pedersen, 1966; Tarnoczy, 1971). Owing to the directionality of the human mouth/head and torso (Dunn and Farnsworth, 1939), the high frequencies will mainly be radiated in the frontal direction while the diffuse field will represent a spatial integration of radiation in all directions. Furthermore, most room surfaces have a greater absorptivity for high-frequency sounds than for low-frequency sounds, so reverberant or diffuse sound fields will tend to have relatively weaker high-frequency components when compared to anechoic sound fields. Consequently, the relative high-frequency content will be higher in the frontal/anechoic conditions. Nonetheless, the differences in LTASS between the chorus studies and most of the other studies are not large and do not affect within-study comparisons of languages. Notable differences may be obtained when speech is recorded with an angle of incidence substantially different from 0° (Studebaker, 1985). Cornelisse *et al.* (1991) recorded speech at the ear of the talker and Stelmachovicz *et al.* (1993) recorded the speech of parents at the ear position of a childwith various postural positions (e. g., sitting adjacent, hip,

cradle). These last two studies reflect (intentionally) head diffraction and body baffle effects as well as variations resulting from the directionality of speech.

In the present study, the recording microphone was placed at 45° incidence to the talker's mouth axis and at a distance of 20 cm. At this close distance, which was selected to optimize the signal-to-noise ratio and to minimize the effect of any reverberant field present, it was considered undesirable to position the microphone directly in front of the mouth because of breath noises from plosive sounds. The use of 45°, compared with 0°, has been shown to result in a relative reduction of about 2 dB at frequencies from 1000 to 5000 Hz and slightly more at 8000 Hz (Studebaker, 1985). Our choice of incidence could be considered to be a compromise between 0°, which maximizes the relative high-frequency content, and reverberant conditions or other angles, which result in varying amounts of reduction in high frequencies.

The LTASS will undoubtably be influenced by the type of speech material that is analyzed. For example, the accuracy with which the Articulation Index (AI) predicts speech recognition test scores will be optimized by using the LTASS of the material in that test (Studebaker and Sherbecoe, 1992). The LTASS of a list of words or nonsense syllables may be very different from the LTASS of running speech especially if the word or syllable list contains frequent repetition of a few phonemes. (A striking example is shown in Fig. 2 of Byrne, 1986 which presents the LTASS of a nonsense syllable test.) Our interest, however, in common with that of the above cited authors, was not in the LTASS of any specific material but rather in deriving a LTASS that would be representative of speech encountered over a range of everyday situations. It seems from the agreement of the above studies and from comparisons of different materials (Benson and Hirsh, 1953) that the choice of material is not critical provided that it is not grossly unrepresentative phomenically, such as speech passages containing repetition of a few phrases.

Although the differences in measurements of the LTASS are small, it would be desirable to establish a standard LTASS for everyday speech. It would be useful if such a LTASS could be taken as representative of all, or a wide range of, languages for use in hearing aid prescriptive procedures. Some prescription procedures are used with clients who speak languages (e.g., Cantonese) for which LTASS measurements have not, to our knowledge, been published. There is, therefore, a practical issue in deciding whether such procedures should be used, without modification, for clients who listen to a language that differs from the one for which the procedure was developed. In a similar vein, the AI and other predictive procedures have been used with various languages although usually the users have made spectral measurements for the language concerned (Aniansson, 1974; Leijon, 1989).

The present study was designed to examine the feasibility of developing a standard LTASS that would represent a wide range of languages or, alternatively, to identify any significant differences that may exist among languages. The research strategy was to record and analyze samples of a wide variety of languages and dialects, according to a standardized protocol. Although it is established that the LTASS varies with vocal effort (Tarnoczy, 1971; Pearsons *et al.*, 1977), the present study was confined to "normal" vocal effort as it seems unlikely that variations with vocal effort would be language dependent. In addition to the LTASS, as defined by its rms levels, it was of interest to consider the dynamic range of speech across frequencies. This was accomplished by determining, at three frequencies, the levels exceeded various percentages of time. This type of information, also presented by Dunn and White (1940), is relevant to predicting understanding of speech and to evaluating the requirements of amplification systems. Dynamic range metrics, such as peak to rms differences, differ according to speech materials (Studebaker and Sherbecoe, 1992) and other factors. It seems possible that dynamic range may also differ across languages but we are not aware of any information on this point. It is well established (Tarnoczy, 1971; Byrne, 1977; Pearsons *et al.*, 1977; Cox and Moore, 1988) that the LTASS differs for men and women. The main difference is in the 100–200 Hz frequency range and reflects the generally lower fundamental frequency range of male voices. Although this difference should apply across languages, and there is evidence that it does (Tarnoczy, 1971), there may be other more subtle differences that could be language or dialect dependent. The present study, therefore, includes separate analyses of data for male and female voices.

TABLE I. Speech sample information.

Language	Country	No. of talkers	Investigators
English	England	32	Bamford, Wilbraham
	Australia	30	Byrne, Dillon, Tran
	New Zealand	21	Powell
	U. S. A., Memphis	22	Cox, Alexander
	U. S. A., Columbus	21	Oyer, Lambert
Swedish	Sweden, Stockholm	22	Hagerman
	Sweden, Linkoping	20	Arlinger
Danish	Denmark	20	Westerman, Ludvigsen
German	Germany	27	Kiessling
French	Canada	20	Hetu
Japanese	Japan	27	Nakanishi
Cantonese	Hong Kong	25	Kei
Mandarin	Hong Kong	21	Lui
Russian	Russia	21	Tavartkiladze, Frolenkov
Welsh	Wales	23	Stephens, Meredith
Singhalese	Wales	21	Stephens, Sirimanna
Vietnamese	Australia	19	Byrne, Dillon, Tran
Arabic	Egypt	20	Kotby, Nasser, El Kholy

I. METHODS

A. Overview

Eighteen speech samples were recorded in 13 countries (15 locations) using standard sets of equipment and a stan-

dard protocol. The recordings were then analyzed at the National Acoustic Laboratories (Australia). The recordings represented 13 languages including English as spoken in four countries. Table I lists the languages and countries represented, the numbers of talkers in each sample, and the investigators who made each recording. Most languages were recorded in the countries of their origin, the exceptions being Singhalese and Vietnamese which were recorded in Wales and Australia, respectively.

B. Talkers

With one exception, each recording contained at least ten male and ten female talkers aged between 15 and 60 years. (There were only nine female Vietnamese talkers.) All talkers spoke the language concerned as their first language and none had any obvious speech defects. No other selection criteria were employed.

C. Recording equipment

Eight sets of recording equipment were assembled. These consisted of a high quality (but inexpensive) cassette tape recorder deck (Technics RS-B105) and a custom-made microphone unit based on a Knowles EA 1934 miniature microphone. A set of equipment and a tape was sent to each investigator and was returned with the recording. During analysis, each recording was replayed on the equipment on which it had been recorded. The frequency response of each recording and playback system was relatively flat and corrections were made for the minor discrepancies that existed (see Sec. I F).

D. Speech material

A passage was selected from a story book on the basis that it was relatively easy to read and did not involve excessive repetition. (Most material would meet this latter criterion except some nursery rhymes.) This "standard" passage was used for all recordings of English but other material, meeting the same criteria, were used for other languages. The material took about 90 s for most talkers to read and provided more than the required 64 s of speech for all talkers.

E. Recording procedure

Whenever possible, recordings were made in an anechoic room at least 3 m wide by 3 m long by 2 m high. However, this criterion was relaxed in several instances because such a facility was unavailable. The talker read the material which was enlarged and placed on a chart at least 1 m in front of him or her. Thus the talker was able to look straight ahead throughout the recording and the procedure avoided any possibility of reflections from material held in the talker's hands. The recording microphone was on a stand or suspended (i. e., not on a table or other reflecting surface) in front of the talker, 20 cm from and in the same horizontal plane as the mouth and at an azimuth of 45° incidence, relative to the axis of the mouth.

The talker was instructed to read aloud at a normal speed and level. Before recording, each talker read the passage silently and then read it aloud at least once to ensure reasonable fluency. The talker was instructed that it did not matter if there were some mistakes and that he or she should keep on reading rather than stopping to correct any errors. The practice reading was also used to set the tape recorder to an appropriate recording level (adequate but with minimal overloading) and this setting was noted so that absolute levels could be calculated later. (Although not calibrated, the recorder volume control was large and well marked and thereby permitted settings to be reproduced accurately.)

F. Analysis procedures

Each recording of speech was analyzed, using a Bruel & Kjaer 2131 analyzer coupled to a Tektronics computer, to derive overall and third-octave band rms levels, averaged over 64 s of signal. Absolute levels were derived by comparing the output intensity to that of a standard pure-tone calibrator which was used to record an 84 dB SPL tone on the tape and was replayed.

The frequency response of each microphone and tape recorder was analyzed at the center of each 1/3 octave frequency from 100 to 10 000 Hz. As the variation between the correction figures for each set of equipment relative to the average correction figure was less than 1.2 dB, and for most frequencies was less than 0.5 dB, a common set of correction figures was used for all recordings. Values for 63, 80, 12 000, and 16 000 Hz were obtained by extrapolation, and the results at these frequencies should thus be treated with some caution. (In fact, the LTASS values for 63 Hz are questionable because they may also be influenced by noise.) The largest correction figure used was +7.5 dB, for 16 000 Hz. Within the range 200–8000 Hz, no correction figure exceeded 2.0 dB.

The dynamic range of the speech was assessed by the following procedure. Ten talkers (five female and five male) from each sample were randomly chosen for analysis. The tape recorder output was input to four Bruel & Kjaer 2231 sound level meters (SLMs), which had installed the DZ7101 statistics module. One of the sound level meters measured the broadband signal and the other three were connected to Bruel & Kjaer 1625 1/3 octave filter sets. These were set to 400, 1000, and 4000 Hz. The SLMs were set to detect the rms envelope of the input signal using a "Fast" time constant (conforming to IEC 651). This time constant is nominally 125 ms. The SLMs were programmed to measure for 64 s and the following parameters were recorded from each of the four sound level meters at the end of the analysis period. "Peak" refers to the highest instantaneous signal level measured during each passage. "Max" refers to the highest level of the rms envelope measured during each passage. "L_{eq} refers to the long-term rms value during the entire analysis time. "L1" refers to the envelope value exceeded 1% of the time. $L10$, $L50$, $L90$, and $L99$, similarly refer to the envelope values exceeded 10%, 50%, 90%, and 99% of the time, respectively. The SLM sampled the envelope at intervals of 31.25 ms.

G. Statistical treatment

The long-term equivalent 1/3 octave and overall levels obtained from the spectrum analysis were processed in two ways. The first concentrated on examining differences in spectral shape, while the second concentrated on examining differences in overall level.

One-third octave levels for each talker were corrected by the measured frequency response of the tape recorder and microphone. These values were normalized so that the long-term overall (linear frequency weighting) level was 70 dB. Two-way analysis of variance (ANOVA) was performed with "Sample" as a between-groups variable and "Frequency" as a within-groups variable. ("Sample" is used in preference to "country," "language," or "location" because, in some instances, two languages were recorded in the same location and, in other instances, the same language was recorded in two or more countries or locations.) Separate analyses were performed for the male and female talkers.

To enable a ready comparison among samples, the grand average spectrum was calculated by averaging, separately for males and females, the mean spectrum for all samples excluding Arabic (see later). The mean for each sample, males and females separately, was then plotted against the grand average spectrum.

The significance of deviations from the grand average spectrum was assessed as follows. The grand average spectral levels were subtracted from the normalized 1/3 octave levels at each frequency for each talker. At each frequency, a t test was then used to compare the mean difference for each sample (Arabic excluded) with the null hypothesis of zero difference. Because there were 850 such tests (25 frequencies by 17 samples by two sexes), we would expect 42 to be significant by chance alone if the usual significance level of 0.05 was adopted. For this reason, Figs. 1–4 show only which deviations from the grand average are "significant" with $p<0.01$.

The overall levels in dB SPL with a flat frequency weighting were analyzed in a two-way ANOVA with sex and sample as between-group variables.

H. Analysis of Arabic

The Arabic speech, which became available later than the other samples, received a more limited analysis. The LTASS was determined by the procedures described above but dynamic range measurements were not undertaken. Arabic was not included in the statistical treatment (ANOVAs) or in determining the "universal" LTASS to be presented later. The significance of differences between Arabic and the universal LTASS was assessed by t tests.

II. RESULTS

A. Shape of the speech spectrum

For both the male and female talkers, the major effects of sample and frequency, and the interaction between, them were all significant with $p<0.000\,001$. Figures 1–4 show the male and the female LTASS for each sample, except Arabic. Also shown is the average LTASS for all samples combined. Separate male and female values are shown for frequencies below 200 Hz but, for the rest of the frequency range, the average is for males and females combined. Symbols shown in bold are those deviations from the average that are significant at the 0.01 level and filled symbols show deviations that are significant at the 0.001 level.

Considering Figs. 1–4, it is clear that the number of "significant" deviations (204) is far more than would be expected by chance alone (eight). In view of the numerous comparisons, we suggest that deviations should only be considered to be truly significant when they are at the 0.001 level, or when deviations at the 0.01 level occur in two or more adjacent frequency bands.

The values for the grand average spectra (included in Figs. 1–4) for males, and females, are shown in Table II. The third column shows a combined long-term average spectrum. For frequencies up to and including 160 Hz, it is equal to the male spectrum. For higher frequencies, it is equal to the average of the values for males and females. As will be discussed later, this is recommended as an appropriate universal spectrum for many, but not all, purposes.

B. Overall levels

The ANOVA (with non-normalized data) showed that sample was highly significant ($p<0.000\,001$) but that sex was not significant at the 0.05 level. The average value for males was 71.8 dB SPL, while that for females was 71.5 dB SPL. The average values for each sample (both sexes combined) ranged from 67.8 for Vietnamese to 75.2 dB for Mandarin. The distribution of overall levels for all talkers is shown in Fig. 5.

C. Dynamic range

All values for each talker were normalized by subtracting that talker's L_{eq} value for the respective band. The resulting relative levels are shown in Fig. 6. The only major variations between the samples occurred for the $L90$ and $L99$ percentile levels. These may not be indicative of genuine differences between the languages and dialects because these levels are presumably affected by the amount of background noise present during pauses in the continuous discourse.

A four-factor ANOVA was performed on the data with the $L90$ and $L99$ data excluded. Between-groups factor were sample and sex, and repeated measures factors were band and percentile. The main effects are of no interest, and of the 11 interaction effects the following were significant: sex × band, sample × percentile, sex × percentile, band × percentile, sample × sex × percentile, sample × band × percentile, sex × band × percentile, and sample × sex × band × percentile. With the exception of sample × sex × percentile ($p=0.008$), all of these were significant with $p<0.000\,2$. Despite the extremely high level of significance, the effects were generally only a few decibels in magnitude, and the statistical significance arose from the large number of talkers and observations. Figure 7 shows all the percentile levels as a function of band and sex.

D. Individual talker differences

The spectra for individual talkers showed substantial variations. This was true for all samples although there was

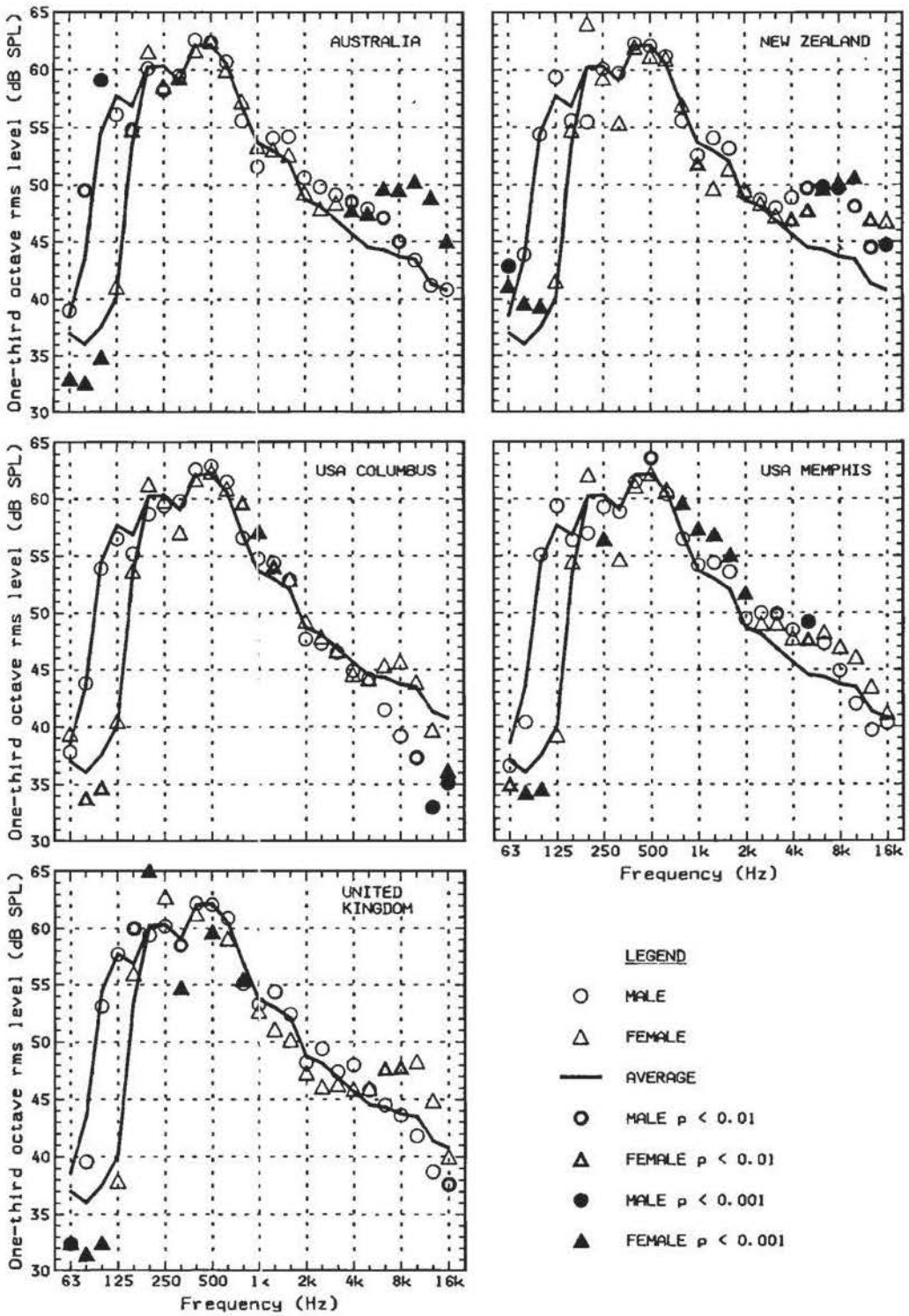

FIG. 1. Male and female long-term average speech spectrum (LTASS) values for five samples of English. Solid line shows LTASS average across 17 speech samples (all samples except Arabic), males and females separately for frequencies below 160 Hz, combined for higher frequencies.

no statistical examination of whether variability interacted with sample. For all samples combined, Fig. 8 shows the standard deviation of individual variations from the mean value at each frequency for males and for females. The analysis is based on data which had been normalized to a 70 dB overall level for each talker. The deviation values are similar for both sexes and at all frequencies from 630 to 4000 Hz. Variability shows a small but consistent increase,

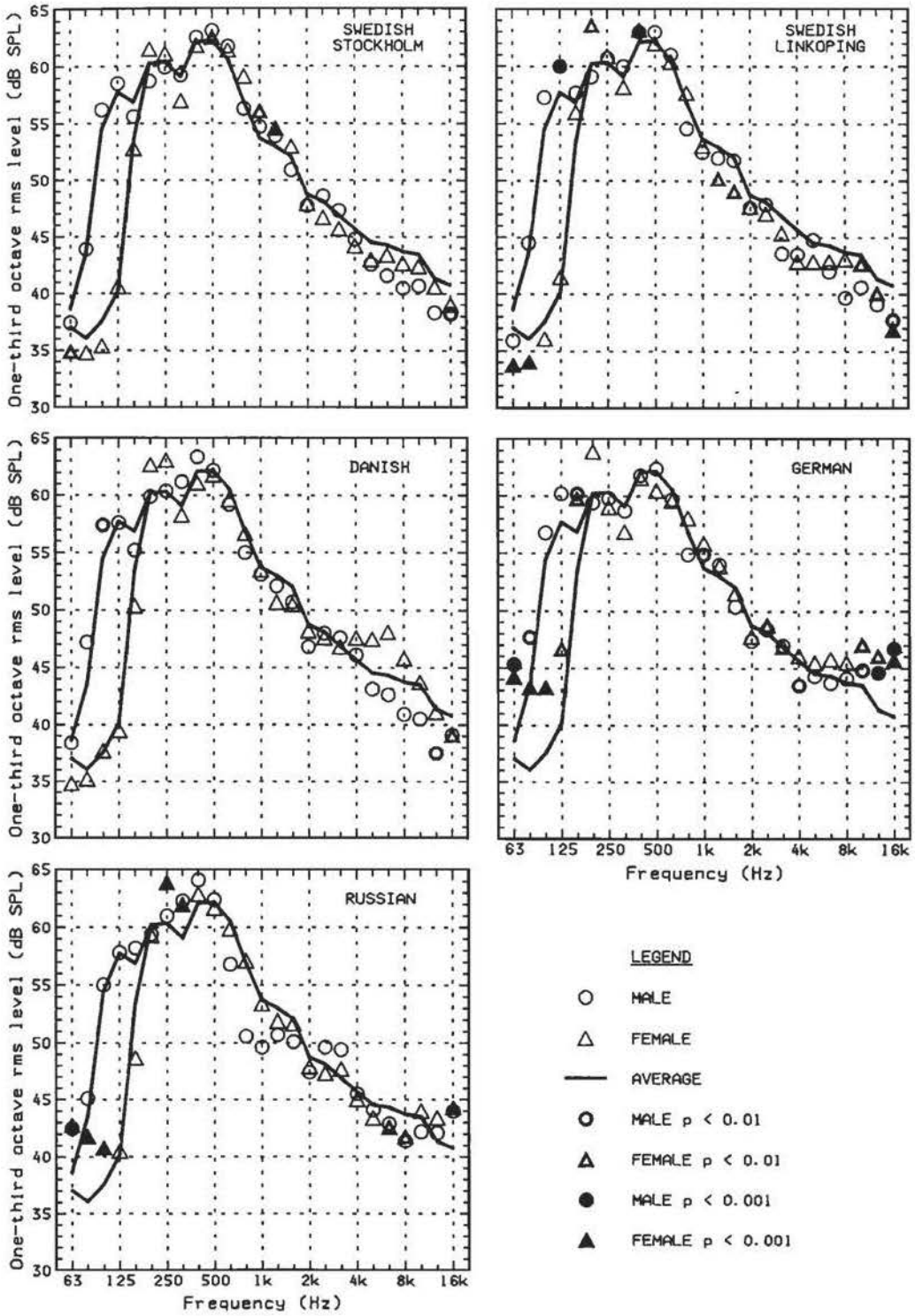

FIG. 2. Male and female LTASS values for Swedish (two samples), Danish, German, and Russian. Solid line shows LTASS average across 17 speech samples.

for both sexes, for bands above 4000 Hz. The only substantial increases in variability are in the bands 80 and 100 Hz for males and 125 and 160 Hz for females.

E. LTASS for Arabic

The LTASS values (dB) for Arabic, normalized to 70 dB overall level (linear), are shown in Fig. 9.

III. DISCUSSION

The overall finding of this study is that the LTASS is very similar over the wide range of languages that were analyzed. Indeed, there is no single language or group of languages which could be regarded as being markedly different from the others. Therefore, it is feasible to propose a univer-

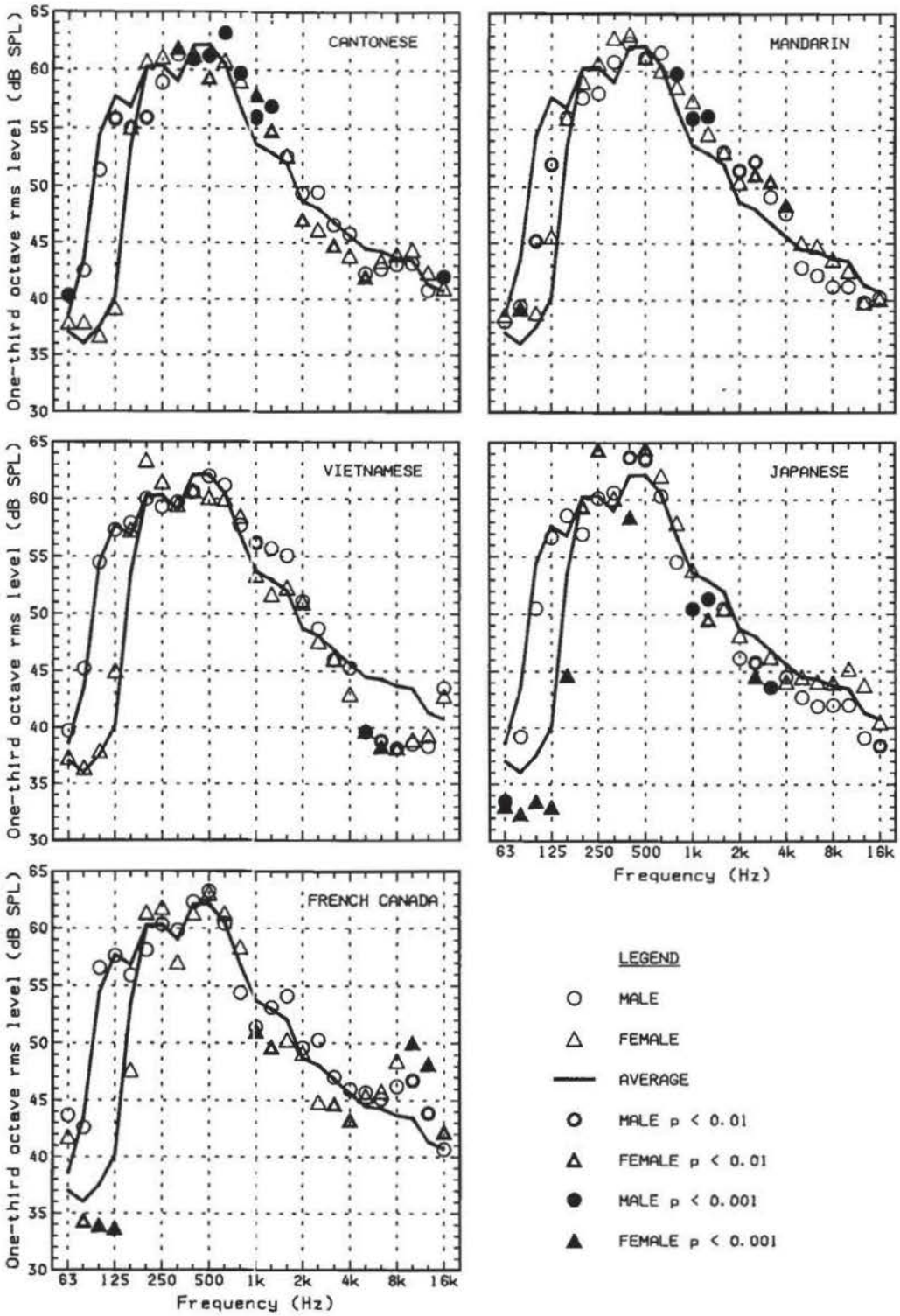

FIG. 3. Male and female LTASS values for Cantonese, Mandarin, Vietnamese, Japanese, and French. Solid line shows LTASS average across 17 speech samples.

sal LTASS that would be applicable to most (possibly all) languages and would be sufficiently precise for many purposes. Nonetheless, there are small but (statistically) significant differences among languages and more substantial differences, at the low frequencies, between male and female talkers.

A. Male/female differences

Considering first the comparison between males and females, the most notable feature is that their spectra are virtually identical over the frequency range from 250 to 5000 Hz. Within this range, the normalized male and female lev-

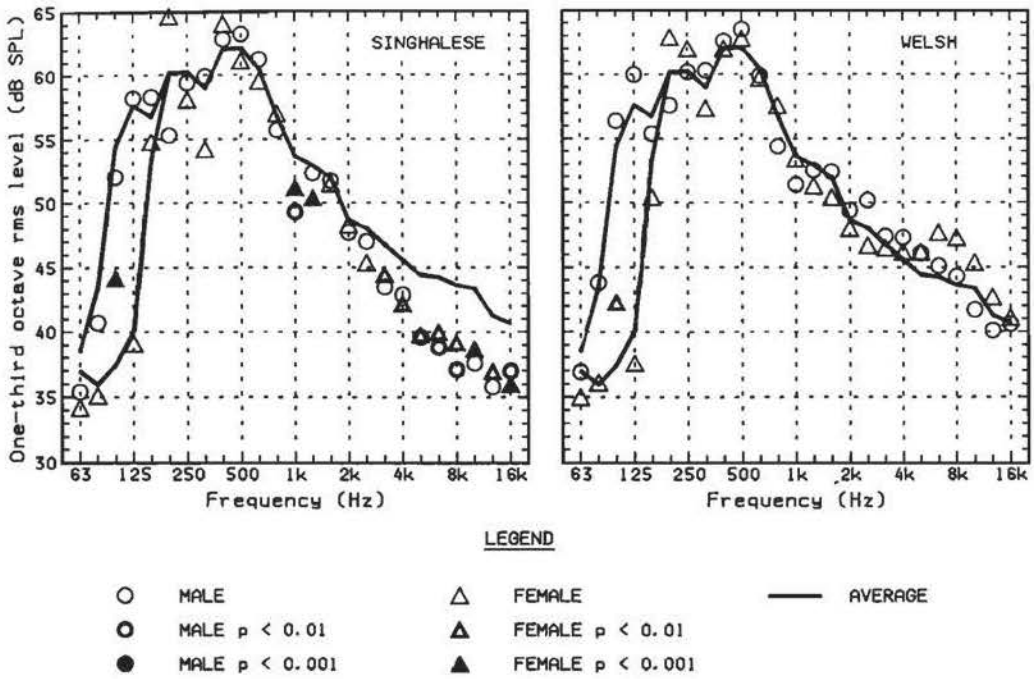

FIG. 4. Male and female LTASS values for Singhalese and Welsh. Solid line shows LTASS average across 17 speech samples.

TABLE II. Male, Female, and combined speech spectra, normalized for 70 dB SPL overall level and averaged across samples (excluding Arabic). Combined is equal to male for frequencies up to 160 Hz; it is average of male and females for other frequencies. The spectrum (combined male and female) recommended by Cox and Moore (1988) is also shown.

Frequency (Hz)	Male	Female	Combined	Cox and Moore (1988)
63	38.6	37.0	38.6	...
80	43.5	36.0	43.5	...
100	54.4	37.5	54.4	...
125	57.7	40.1	57.7	...
160	56.8	53.4	56.8	...
200	58.2	62.2	60.2	...
250	59.7	60.9	60.3	60.0
315	60.0	58.1	59.0	57.0
400	62.4	61.7	62.1	61.0
500	62.6	61.7	62.1	62.0
630	60.6	60.4	60.5	59.0
800	55.7	58.0	56.8	56.5
1000	53.1	54.3	53.7	55.0
1250	53.7	52.3	53.0	54.5
1600	52.3	51.7	52.0	52.0
2000	48.7	48.8	48.7	49.0
2500	48.9	47.3	48.1	48.0
3150	47.0	46.7	46.8	46.5
4000	46.0	45.3	45.6	46.0
5000	44.4	44.6	44.5	44.0
6300	43.3	45.2	44.3	45.5
8000	42.4	44.9	43.7	...
10 000	41.9	45.0	43.4	...
12 500	39.8	42.8	41.3	...
16 000	40.4	41.1	40.7	...

els, averaged over all languages, agree within 2 dB at all third-octave frequencies except 800 Hz, where the difference is 2.3 dB. (The non-normalized values agree almost as closely as there was little difference between the average overall male and female speech levels.) For frequencies of 160 Hz and below, male levels greatly exceeded female levels undoubtably because of the difference in the fundamental frequency ranges. These findings are consistent across languages and consistent with previous research (Benson and Hirsh, 1953; Tarnoczy and Fant, 1964; Tarnoczy, 1971; Niemoller *et al.*, 1974; Byrne, 1977; Pearsons *et al.*, 1977; Cox and Moore, 1988).

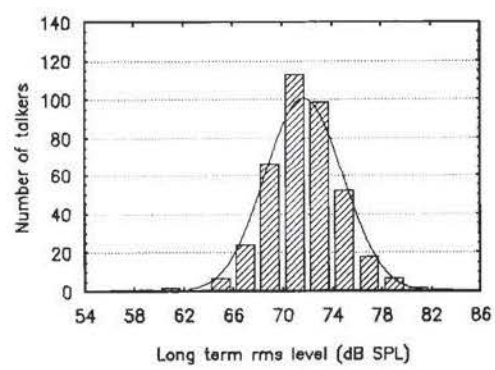

FIG. 5. Distribution of overall rms levels (measured at 20 cm from mouth). Curve shows normal distribution fitted to data.

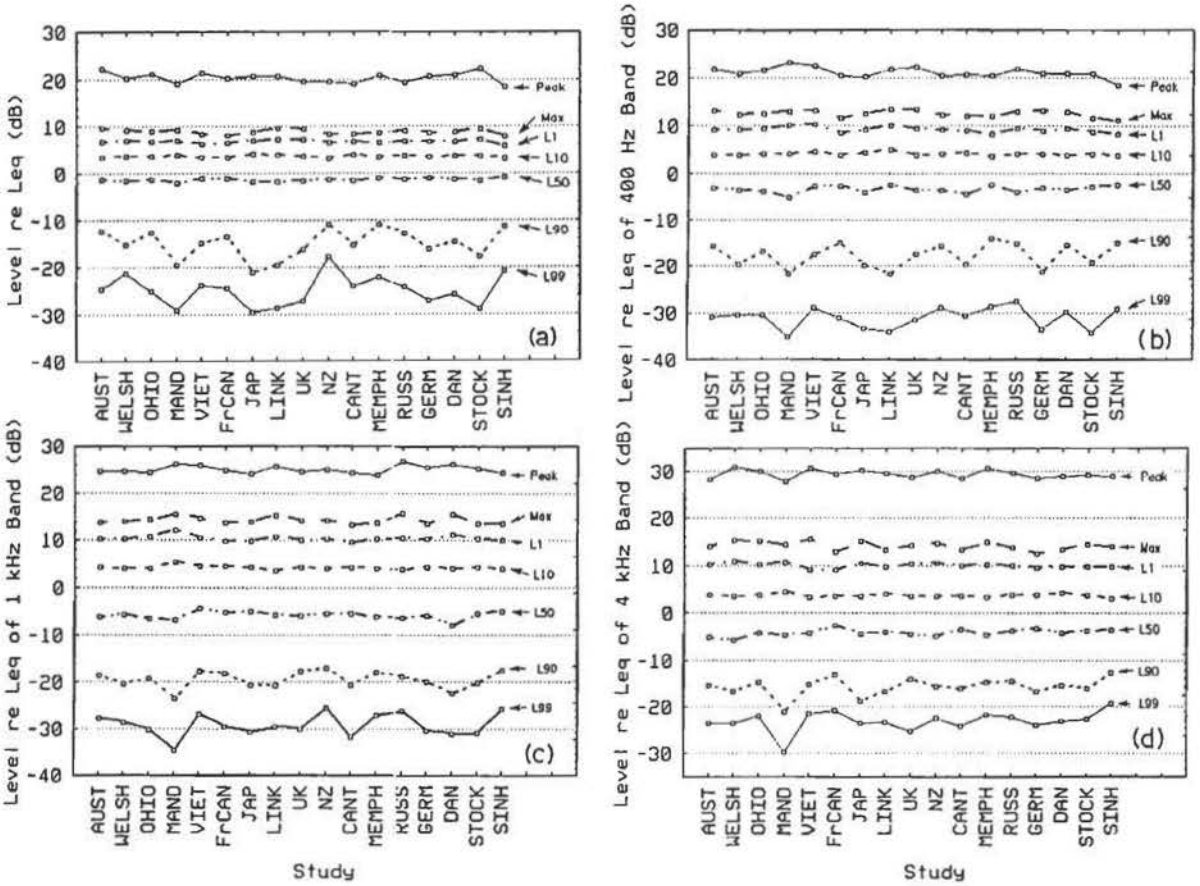

FIG. 6. (a) Peak, maximum, and percentile levels of the broadband signal (relative to L_{eq}) for each sample. Peak, maximum, and percentile levels of the (b) 400, (c) 1000, and (d) 4000 Hz 400 1/3 octave bands (relative to the L_{eq} for that band) for each sample.

For 6300 Hz and higher frequencies, female levels exceeded male levels. This was a consistent finding; when averaged across the frequencies from 6300 to 12 000 Hz inclusive, the average level for females exceeded that for males for every country although only marginally so in some instances. This is shown in Figs. 1–4. The average difference between males and females, in this range, was 2.6 dB, with differences ranging from 0.3 to 5.9 dB. The same trend occurred in the Arabic data (Fig. 9) in that the female values exceeded the male values by an average of 2.9 dB.

Some previous studies have found a higher overall level, typically 2–3 dB, for male than for female voices (Benson and Hirsh, 1953; Byrne, 1977; Pearsons *et al.*, 1977). Our data do not show a significant difference although the small

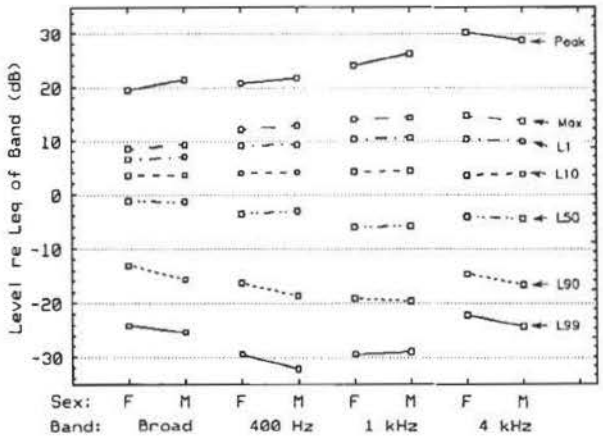

FIG. 7. Percentile levels (relative to band L_{eq}) versus talker sex and measurement band.

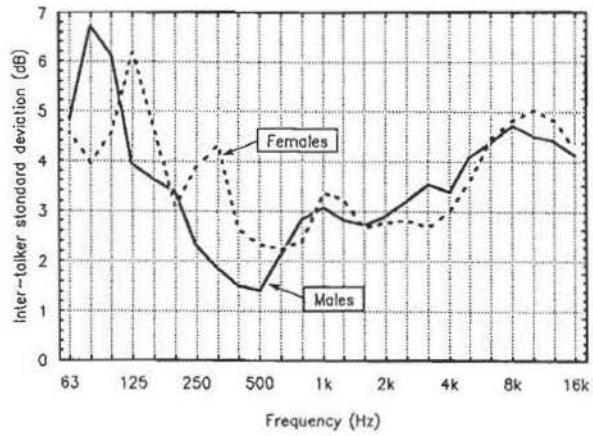

FIG. 8. Individual variability in speech levels averaged across 17 samples (all samples except Arabic) normalized for L_{eq}.

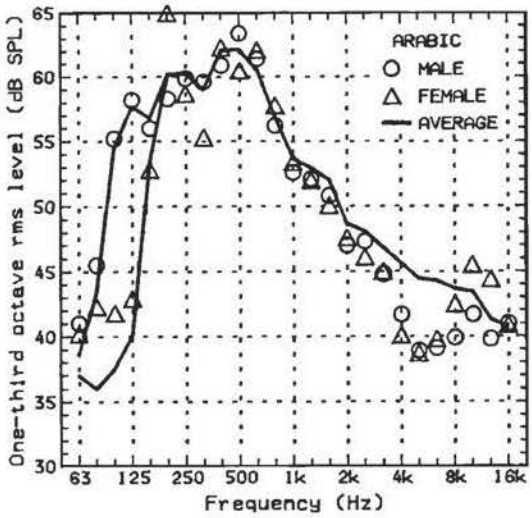

FIG. 9. Male and female long-term average speech spectrum for Arabic. Solid line shows LTASS averaged across the other 17 samples (i.e., excluding Arabic).

mean difference that exists (0.3 dB) is in the predicted direction. Pavlovic *et al.* (1991) have also reported no significant difference for sex. The discrepancy in the findings of different studies may be influenced by differences in vocal effort as Pearsons *et al.* found that the male/female difference in overall level increased with increasing vocal effort. Also, any intrinsic male/female differences could possibly be reduced by instructions if they discouraged the use of particularly soft or loud voices.

There appears to be some male/female difference with respect to the degree of individual talker variability in the lower frequency bands. This affects mainly the frequency bands somewhat below the typical male and female fundamental frequencies and is probably related to differences in the fundamental frequency range of individual talkers.

B. Language/dialect differences

Individual languages showed many statistically significant variations from the average values. However, most of the variations are less than 3 dB in magnitude, or occur outside the range of 200 to 6300 Hz, or occur only in one isolated 1/3 octave band. If we examine only those statistically significant variations which are larger than 3 dB, and which occupy at least two adjacent bands within the range 200 to 6300 Hz inclusive, we find only the following variations from a "universal speech spectrum:"

(1) New Zealand male and female speech is high by 3–6 dB from 5000 Hz and above;
(2) Vietnamese male and female speech is low by 5–6 dB from 5000 Hz and above;
(3) Arabic male and female speech is low by 5 dB at frequencies around 5000 Hz;
(4) Japanese male speech is low by 3 dB around 2500 Hz;
(5) Cantonese male speech is high by 3–4 dB from 630 to 1250 Hz;
(6) Australian female speech is high by 3–4 dB from 5000 Hz and above;
(7) Memphis female speech is high by 3–4 dB from 1250 to 2000 Hz;
(8) Russian female speech is high by 3–4 dB at 250 and 315 Hz;
(9) Mandarin female speech is high by 3–4 dB from 2500 to 4000 Hz; and
(10) Singhalese female speech is low by 3–5 dB from 4000 Hz and above.

Explanations for the above variations are not obvious. It may be that some or all of them occur because different languages or dialects use somewhat different vowels or use the same vowels or other sounds, such as "s," with different frequencies of occurrence. Previous authors have sometimes suggested explanations of this type for the differences they found among the LTASS of different languages (Tarnoczy and Fant, 1964; Harmegnies and Landercy, 1985) or dialects (Cox and Moore, 1988). However, our data show that when languages deviated from the average, they often did so for only one sex. Therefore, any factors which explain the deviations must interact with sex differences. The difficulty of finding credible explanations for differences in the LTASS may be illustrated by comparing the five samples of English (Fig. 1) with respect to sex differences. In the frequency range above 5000 Hz, male/female level differences occur for three samples but not for the other two samples. It is difficult to understand how this sex difference could occur in Australian speech but not in New Zealand speech, or how it could occur in one American sample but not in the other. It is clear that it would be a major undertaking to attempt to relate differences in languages or dialects to differences in the LTASS. Such an investigation was beyond the scope of the present study.

An obvious question is whether some or all of the small differences found could be explained by differences in the recording techniques used at different locations. Against this explanation is the point just mentioned, namely that differences from the LTASS often occurred for only one of the sexes. Furthermore, variations tend to occur over several adjacent third-octave bands. It therefore seems unlikely that the variations are caused by room resonances, which tend to be more localized in frequency. We note also that in the instances where two languages were recorded in the same locality, they do not show similar deviations. Thus it seems highly unlikely that differences in recording technique could explain the differences noted above. (Possible exceptions could be the New Zealand and Arabic samples as both sexes show the same trend and there are no other recordings from the same locality for comparison.)

From considering the above noted variations and others occurring at frequencies below 200 or above 6300 Hz, there appears to be no systematic separation between the English versus the non-English languages, or between the nontonal versus the tonal languages (Cantonese, Mandarin, Vietnamese). This last finding agrees with that of McCullough *et al.* (1993) who found no difference between the LTASS of English and Mandarin.

C. Effects of normalization/analysis methods

Most of the above deviations occur in the high-frequency or midfrequency region. However, the nature of the deviations observed will depend partly on the choice of analysis method. The normalization process ensured that the overall level of all talkers was set to 70 dB SPL. Because the overall level is dominated by the more intense low-frequency bands, there is less chance of observing differences among talkers in this region. The alternative of normalizing at one specific frequency would cause the same problem to a greater degree. The use of non-normalized levels would make the detection of differences in shape dependent on the range of overall levels allowed by each experimenter because variation of overall level would add to the variance among talkers at each frequency.

The significance of the major effect of "sample" within the analysis of variance of spectral shape is at first surprising because the data had been normalized so that all talkers had an overall level of 70 dB SPL. The effect is readily explained by variations in the high-frequency region. Because the speech spectrum is weighted towards the low frequencies, the overall level, based on a power addition of the individual bands, is little affected by the level of the lower intensity, high-frequency components. The ANOVA statistic, however, sums the decibel value of all bands linearly, so variations in the high-frequency level affect the major effect of sample just as much as do variations in the low-frequency levels.

D. Dynamic range

The dynamic range measurements, like the LTASS measurements, show that all languages are similar despite there being a number of statistically significant, but small, differences. Essentially, the $L50$, $L10$, $L1$, Max, and Peak values are equivalent across all samples (Fig. 6) and for males and females (Fig. 7). There is one difference between our data and those of other studies in that our data show a 10-dB difference between the $L1$ and L_{eq} values in contrast to the widely accepted value of 12 dB (ANSI, 1959), although only slightly different from the value of 11 dB shown by the data of Cox *et al.* (1988). Another minor discrepancy is that our data do not show the $L1/L_{eq}$ difference to increase with frequency, as is shown by some other data (e.g., Fletcher and Galt, 1950).

E. Overall level

The measured overall speech levels of our subjects averaged about 72 dB SPL for both male and female talkers. As the recording microphone was only 20 cm from the mouth, this level would correspond to 58 dB SPL at a distance of 1 m, assuming no reverberation. This level is about 5 dB less than is usually reported for conversational speech (Pearsons *et al.*, 1977). We suggest that, despite instructions to speak "normally," many talkers tend to speak at a low level when a microphone is placed close to them. This may be analogous to the natural tendency to adjust voice levels according to the distance from listeners. In similar vein, it has been shown repeatedly (see Byrne, 1983 for review) that the typical speech input received by moderately impaired hearing aid wearers is about 70 dB SPL which requires greater than normal vocal effort. We believe that no significance should be attached to the overall levels found in a study such as ours except to note that differences in vocal effort affect the LTASS. Greater vocal effort will result in an increase in the relative mid- to high-frequency content of speech, although the differences between soft and average speech are small (Pearsons *et al.*, 1977; Kiukaanniemi *et al.*, 1982).

F. Universal LTASS

The similarity of the LTASS across samples demonstrates that it is reasonable to propose a universal LTASS which should be satisfactory for many purposes and applicable to most, if not all, languages. Such a LTASS should be suitable for hearing aid prescription procedures and for the Articulation Index, regardless of language. The comparability of speech dynamics across languages also supports the idea that the application of such procedures to different languages is not complicated by acoustical differences in conversational speech. Of course, there may well be other complications, such as possible differences in frequency importance functions across languages, which would need to be investigated before applying procedures like the Articulation Index universally. Our recommended universal LTASS is very similar to the recommendation of Cox and Moore (1988) (see Table II) but it has the advantages of having been derived from a larger total data set and of representing a range of languages. It may also be more realistic because it is based on a 45° rather than 0° angle of incidence.

As mentioned earlier, substantially different speech spectra have been obtained for measurements made at different angles of incidence with respect to the talker and when influenced by variations in head/body diffraction effects (Cornelisse *et al.*, 1991; Stelmachovicz *et al.*, 1993). Those measurements were prompted by an interest in prescribing amplification for children and the recognition that speech is often presented from directions other than directly in front of the child. In a revised version of a hearing aid prescription procedure, Seewald (1992) opted to use a compromise between the usual (0° incidence) LTASS and one recorded at the ear of the talker. The possible merits of such a choice will not be considered here but it is mentioned to show that there may be an argument for using different LTASS values for particular purposes. Nonetheless, our universal LTASS could serve as a basis from which any required variations could be made. There could be circumstances where separate male and female spectra would be desirable, namely applications where the very low frequencies are significant. However, a single LTASS is sufficient and, therefore, preferable for the applications considered here, that is, in relation to hearing aid prescription and the AI or similar procedures for predicting speech intelligibility. Our universal LTASS should also be sufficiently precise for a range of more general applications concerning the design or use of speech transmission systems.

Finally, it is clear that the LTASS and the dynamic characteristics of conversational speech are very much dominated by the characteristics of the vocal mechanism. Although different languages use different vowels (formant structures)

and the frequency of occurrence of various phonemes differs, these factors appear to have had only minor effects on the LTASS and dynamic measures of conversational speech. Furthermore, these differences are not even consistent across male and female talkers of the same language.

ACKNOWLEDGMENTS

In addition to the authors, the following participated in the study: John Bamford (University of Manchester), Genevieve Alexander (VA Medical Center, Memphis), and C. Lambert (Ohio State University). We also thank Dr. Margaret W. Skinner for helpful comments on a draft of this manuscript.

Aniansson, G. (**1974**). "Speech discrimination predicted from tone audiometry and articulation index," Acta Oto-Laryngol. Suppl. **320**, 36–43.

ANSI (**1969**). ANSI S3.5-1969, "American National Standard Methods for Calculation of the Articulation Index" (American National Standards Institute, New York).

Banuls-Terol, V. (**1971**). "Weighted average spectrum of human speech: An approach," Proceedings of the 7th ICA, Budapest, pp. 253–256.

Benson, R. W., and Hirsh, I. J. (**1953**). "Some variables in audio spectrometry," J. Acoust. Soc Am. **25**, 499–505.

Berger, K., Hagberg, N. S., and Rane, R. L. (**1977**). *Prescription of Hearing Aids* (Herald, Ohio).

Boothroyd, A. (**1967**). "The discrimination by partially hearing children of frequency distorted speech,"Int. Audiol. **6**, 136–145.

Byrne, D. (**1977**). "The speech spectrum—Some aspects of its significance for hearing aid selection and evaluation," Br. J. Audiol. **11**, 40–46.

Byrne, D. (**1983**). "Theoretical prescriptive approaches to selecting the gain and frequency response of a hearing aid," Monogr. Contemporary Audiol. **4**, (1), 1–40.

Byrnes, D. (**1986**). "Effects of frequency response characteristics on speech discrimination and perceived intelligibility and pleasantness of speech for hearing-impaired listeners," J. Acoust. Soc. Am. **80**, 494–504.

Byrne, D., and Dillon, H. (**1986**). "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," Ear Hear. **7**, 257–265.

Dalsgaard, S. C., and Pedersen, S. B. (**1966**). "The power density spectrum of Danish speech," Transactions of Danish Engineering, No. 4-1966, 1–7.

Cornelisse, L. E., Gagné, J. P., and Seewald, R. C. (**1991**). "Ear level recordings of the long-term average spectrum of speech," Ear Hear. **12**, 47–54.

Cox, R. (**1988**). "The MSU hearing aid prescription procedure," Hear. Instrum. **39**, 6–10.

Cox, R. M., Matesich, J. S., and Moore, J. N. (**1988**). "Distribution of short-term rms levels in conversational speech," J. Acoust. Soc. Am. **84**, 1100–1104.

Cox, R. M., and Moore, J. N. (**1988**). "Composite speech spectrum for hearing aid gain prescriptions," J. Speech Hear. Res. **31**, 102–107.

Dunn, H. K., and Farnsworth, D. W. (**1939**). "Exploration of pressure field around the human head during speech," J. Acoust. Soc. Am. **10**, 184–199.

Dunn, H. K., and White, S. D. (**1940**). "Statistical measurements on conversational speech," J. Acoust. Soc. Am. **11**, 278–289.

Fletcher, H., and Galt, R. H. (**1950**). "The perception of speech and its relation to telephony,"J. Acoust. Soc. Am. **22**, 89–151.

Harmegnies, B., and Landercy, A. (**1985**). "Language features in the long-term average spectrum," Rev. Phonét. Appl. **73–74–75**, 69–79.

Harris, C. M., and Waite, W. M. (**1965**). "Measurements of speech spectra recorded with a close-talking microphone," J. Acoust. Soc. Am. **37**, 926–927.

Kiukaanniemi, H., Soponen, P., and Mattila, P. (**1982**). "Individual differences in the long-term speech spectrum," Folia Phoniatr. **34**, 21–28.

Kiukaanniemi, H. (**1980**). "Speech intelligibility in hearing losses linearly sloping to high frequencies," Acta Univ. Oul. D 57. Opthalmol. Oto-rhino-laryngol. 6. 1–69.

Leijon, A. (**1989**). Optimization of hearing-aid gain and frequency response for cochlear hearing losses, Technical Rep. No. 189, Chalmers University of Technology, Gothenburg, Sweden.

McCullough, J. A., Tu, C., and Lew, H. L. (**1993**). "Speech-spectrum analysis of Mandarin: Implications for hearing-aid fittings in a multi-ethnic society," J. Am. Acad. Audiol. **4**, 50–53.

Niemoller, A. F., McCormick, I., and Miller, J. D. (**1974**). "On the spectrum of spoken English," J. Acoust. Soc. Am. **55**, 461.

Pavlovic, C. V. (**1989**). "Speech spectrum considerations and speech intelligibility predictions in hearing aid evaluations," J. Speech Hear. Disord. **54**, 3–8.

Pavlovic, C. V., Rossi, M., and Espesser, R. (**1991**). "Perceived spectral energy distributions for EUROM.O speech and for some synthetic speech," Proceedings of the XII International Congress on Phonetic Sciences, **5/5**, 418–421.

Pearsons, K. S., Bennett, R. L., and Fidell, S. (**1977**). Speech levels in various noise environment, EPA Rep. No. 600/1-77-025 Environmental Protection Agency, Washington DC.

Rudmose, H. W., Clark, K. C., Carlson, F. D., Eisenstein, J. C., and Walker, R. A. (**1958**). "Voice measurements with an audio-spectrometer," J. Acoust. Soc. Am. **20**, 503–512.

Seewald, R. C. (**1992**). "The desired sensation level method for fitting children: Version 3.0," Hear. J. **45**, 36–41.

Seewald, R. C., Ross, M., and Spiro, M. K. (**1985**). "Selecting amplification characteristics for young hearing-impaired children," Ear Hear. **6**, 48–53.

Skinner, M. W. (**1988**). *Hearing Aid Evaluation* (Prentice-Hall, Englewood Cliffs, NJ).

Stelmachowicz, P. G., Mace, A. L., Kopun, J. G., and Carney, E. (**1993**). "Long-term and short-term characteristics of speech: Implications for hearing aid selection for young children," J. Speech Hear. Res. **36**, 609–620.

Stevens, S. S., Egan, J. P., and Miller, G. A. (**1947**). "Methods of measuring speech spectra," J. Acoust. Soc. Am. **19**, 771–780.

Studebaker, G. A. (**1985**). "Directivity of the human vocal source in the horizontal plane," Ear Hear. **6**, 315–319.

Studebaker, G. A., and Sherbecoe, R. L. (**1992**). "A model for the prediction of average speech recognition performance of normal-hearing and hearing-impaired persons," Lab. Rep. No. 92-02, Hearing Sciences Lab, Memphis State University.

Tarnoczy, von T. (**1971**). "Das durchschnittliche energie-spektrum der sprache (fur sechs sprachen)," Acustica **24**, 57–74.

Tarnoczy, T., and Fant, G. (**1964**). "Some remarks on the average speech spectrum," Q. P. S. R. Rep. No. 4, pp. 13–14, Speech Transmission Laboratory, Stockholm.

Zalewski, J., and Majewski, W. (**1971**). "Polish speech spectrum obtained from superimposed samples and its comparison with spectra of other languages," Proceedings of the 7th ICA, Budapest, pp. 249–252.