Xu, Y. (2019). Prosody, Tone and Intonation. In The Routledge Handbook of Phonetics. W. F. Katz and P. F. Assmann: Routledge, New York. pp. 314-356.

Prosody, Tone, and Intonation

Yi Xu
University College London

1. Introduction:

Prosody refers to all suprasegmental aspects of speech, including pitch, duration, amplitude and voice quality that are used to make lexical and post-lexical contrasts, and to convey paralinguistic meanings. Tone refers to pitch patterns that make lexical, grammatical or morphological contrasts in many languages. Intonation refers to the melodic facet of prosody, although the two terms are sometimes interchangeable. Though often treated as separate research subjects, these three areas are heavily overlapped with each other not only in terms of the phenomena they cover, but also with regard of the underlying mechanisms they share. A common perspective that can link them all is the articulatory-functional principle, which views speech as a system for transmitting communicative meanings with the human articulators (Xu, 2005). From this perspective, lexical, post-lexical, affective and social meanings are all conveyed in parallel, each encoded by its own unique imprints on the properties of various underlying articulatory targets. The articulation process turns these function-loaded targets into the acoustic patterns ultimately perceived by listeners, which consist of not only all the functional contributions, but also all the traces and artefacts of the articulation process itself.

This chapter first outlines the articulatory dynamics underlying the production of pitch, duration and voice quality, which serve as the basis of understanding how different functional components can be realized in parallel. Also introduced are various perturbations by a range of articulatory and aerodynamic effects that accompany the core articulatory process. The functional components are then introduced in terms of their respective contributions to the surface prosody as well as their communicative significance, ranging from making lexical contrast through tone and stress, conveying post-lexical meanings such as focus, sentence modality and boundary marking, to expressing emotional and social meanings through paralinguistic prosody.

The heavy overlap of the communicative functions across the three areas makes it difficult to directly identify the unique contribution of each function to surface prosody. But the sharing of the common mechanisms makes it possible for findings in one area to be relevant to the others, provided that methods are used that can effectively separate the different sources of surface prosody. This chapter is therefore as much about methodology as about the state of the art in prosody research. Effective methodological techniques are illustrated throughout the discussion, and also highlighted at the end of the chapter. Also to be discussed is how the parallel encoding of the communicative functions is related to various prosodic phenomena, including, in particular, prosodic structures and rhythm. Throughout the chapter, various future research directions are suggested.

2. Articulatory mechanisms

Although multiple articulatory mechanisms are at play in the production of prosody, at the core is how the basic dynamics of articulation is used as a coding mechanism. Surrounding this core are sub-processes that affect various aspects of the overall coding dynamics. The effects of these processes are nontrivial, because they make up much of what we see as the prosodic patterns in the acoustic signal. Adequate knowledge about them is therefore crucial for our ability to separate the articulatory artefacts from the underlying targets that encode the communicative functions.

IPR2026-00273
Krisp Technologies EX1025 Page 1

2.1. Basic articulatory dynamics

Speech conveys information by producing rapidly alternating acoustic patterns that represent meaning-carrying categories. In an ideal coding system, the shift between acoustic pattern would be instantaneous, so that surface patterns would directly reflect the underlying coding units without ambiguity. Take as an example fundamental frequency (F_0), the main acoustic correlate of tone and intonation. Instantaneous pattern shifts would look like Figure 1a for a sequence of high and low tones. But the human larynx is incapably of generating such sudden shifts in F_0, because it is a mechanical system subject to inertia. From Figure 1b we can see that the articulated tones look like a series of transitions, each moving gradually toward the underlying register of the corresponding tone, without reaching a plateau.

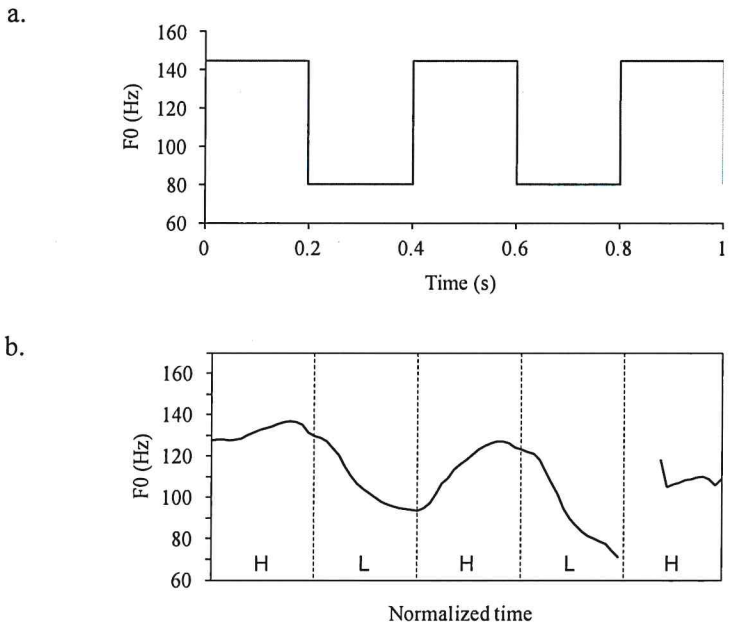


Figure 1. a. A hypothetical F_0 contours of a sequence of high and low level tones with intantaneous transitions. b. Mean F_0 contours of a sequence of high and low tones in Mandarin produced by four native male speakers (data from Xu, 1999).

These transitions seem to be rather slow relative to the normal speech rate of 5-7 syllables/s (Eriksson, 2012; Tiffany, 1980). The relatively slow movement is related to the maximum speed of pitch change (Xu & Sun, 2002). As shown in (1-2), the time needed for pitch change of either direction increases quasi-linearly, and it takes about 100 ms to make even the smallest pitch change (i.e., when d approaches 0).

(1) *Pitch rise:* $t_r = 89.6 + 8.7\,d$

(2) *Pitch drop:* $t_f = 100.04 + 5.8\,d$

where t_r is minimum time needed for a full rise, t_f is the minimum time needed for a full fall, and d is the magnitude of rise or fall in semitones.

From Figure 1b we can see that, for the most part, the transition toward each tonal target occurs *within* the temporal domain of the syllable that carries the tone rather than *between* the syllables. This becomes even clearer in Figure 2, where in each plot only the tones of the second syllable vary across four Mandarin tones: High (H), Rising (R), Low (L) and Falling (F). These tonal variations give the tones of the third syllable very different onsets in F_0. But the four contours in that syllable all converge to a linear shape that is consistent with the underlying form of these tones: high-level, rising and falling, as indicated by the arrows.

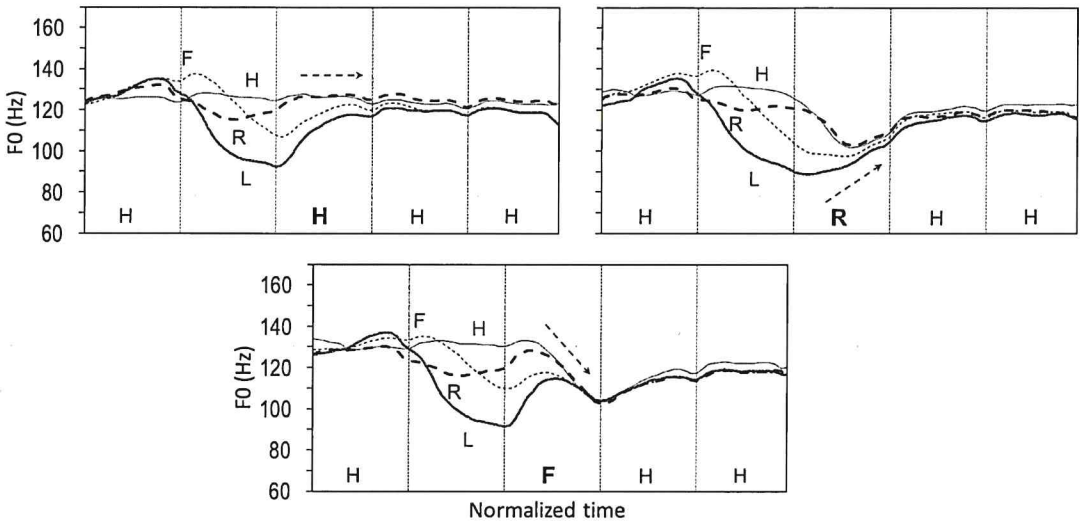


Figure 2. Mean F_0 contours of Mandarin tones in continuous speech, produced by four native male speakers (data from Xu, 1999). In each plot, only the tones of the second syllable vary across four tones.

The convergence of the F_0 contours of the same tone after different preceding tones points to the basic dynamics of tone production: asymptotic approximation of an underlying pitch target during the host syllable, or *syllable-synchronized sequential target approximation* (Xu & Wang, 2001). That is, each tone has an underlying pitch target that can be as simple as a straight line with a specific height and slope, which is approached during articulation in synchrony with the syllable that carries the tone, and adjacent tones are produced by sequentially approaching their respective pitch targets.

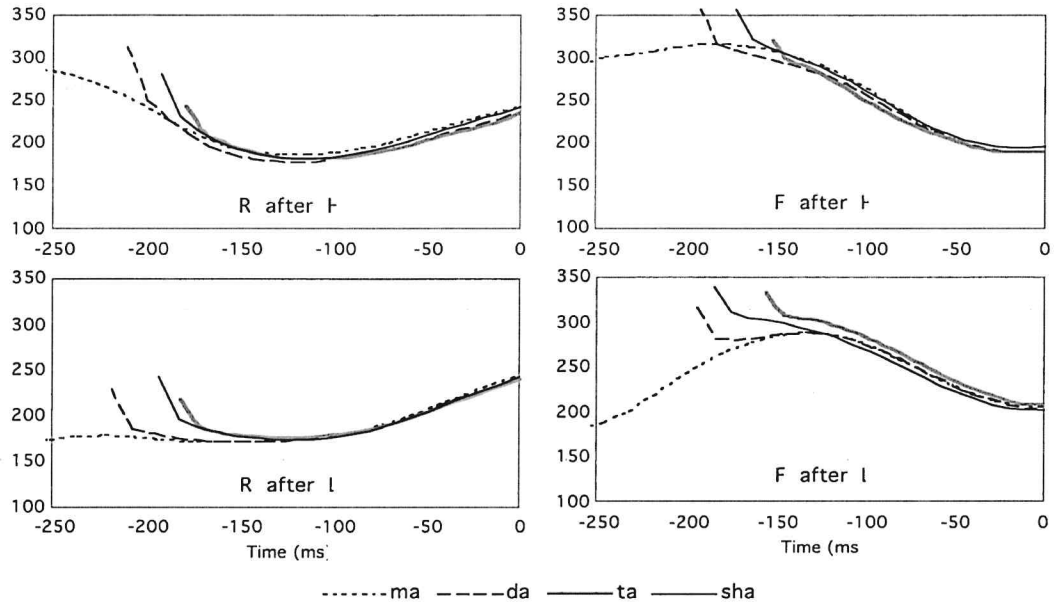
2.2. F_0 perturbation by consonants

When a non-sonorant consonant occurs in a speech utterance, the vibration of the vocal folds is affected in two ways. First, voicing may be stopped, resulting in an interruption of otherwise continuous F_0 contours, or in other words, a *voice break*. Second, F_0 of the surrounding vowels is raised or lowered (Hombert, Ohala & Ewan, 1979), resulting in a *vertical perturbation*. The two effects are very different in nature, but they need to be considered together in order to identify their respective effects on F_0.

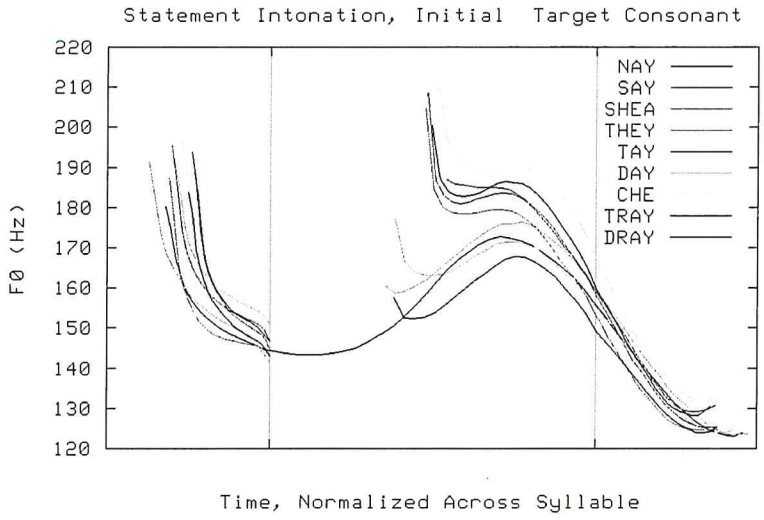
Figure 3 shows F_0 contours of Mandarin (a) (Xu & Xu, 2003) and English (b) phrases (Xu & Wallace, 2004) spoken in carrier sentences. For both languages, the contours of syllables with initial obstruents are displayed together with the contour in a syllable with an initial nasal. As can be seen, an obstruent introduces a gap at the beginning of the syllable. After the gap, F_0 first starts at a level well above that of the nasal, but then drops rapidly towards it. In the Mandarin case, there is almost full convergence with the nasal contour. In the English case the contours are higher than the nasal

contours after a voiceless obstruent, but lower after a voiced obstruent. Thus there are at least two kinds of vertical perturbations, a *sharp upshift* after the voice break followed by a quick drop, and a *voicing effect* that can last till the end of the vowel.

a.



b.



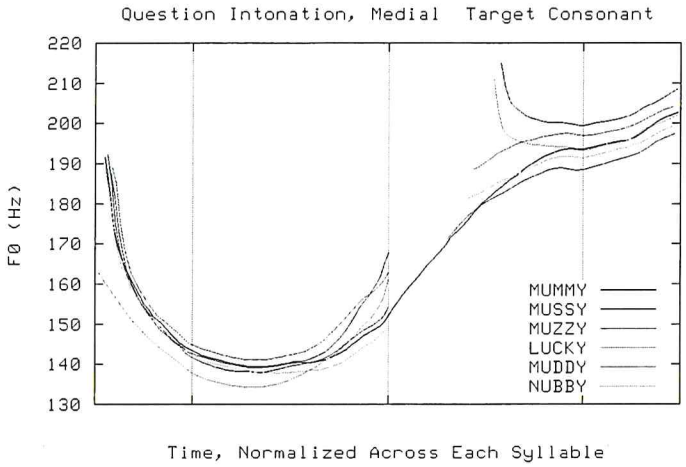


Figure 3. Mean F_0 contours of syllables with a nasal onset (continuous curves) and those with obstruent onsets in a. Mandarin (Xu & Xu, 2003) and b. English (Xu & Wallace, 2004). All contours are time-normalized across the syllable.

As can be seen in Figure 3, the sharp upshift occurs at the release of any obstruent consonants. This means that the effect cannot be due to a high rate of airflow across the glottis at the release of a stop (Hombert et al., 1979). Rather, it is possible that the initial vibration at voice onset after an obstruent involves only the outer (mucosal) layer of the vocal folds, which has a higher natural frequency than the main body of the vocal folds, due to a smaller mass. Mucosal vibration (Giovanni et al., 1999) may also occur at voice offset, where F_0 is often seen to rise in breach of the main tonal or intonational contours, as shown in Figure 4. Here at the end of a declarative utterance, F_0 has dropped to a level of vocal fry. But the vibration suddenly shifts to a new mode at a frequency that is at odds with the general downtrend, and the overall intonation sounds perfectly complete without this final rise. The nature of this possible mucosal vibration (which occurs quite frequently, especially in female voice) and its potential link to the F_0 upshift by an obstruent are both in need of new research.

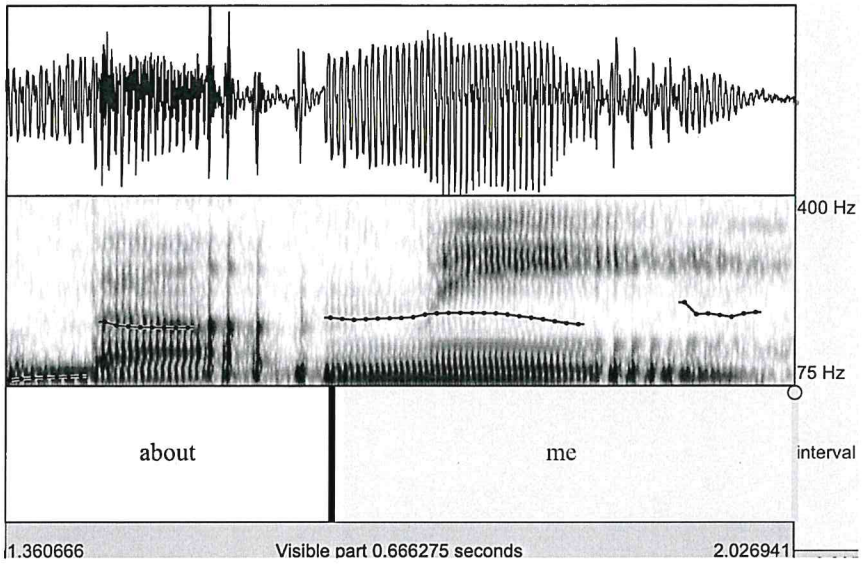


Figure 4. A case of likely mucosal vibration at voice offset. The words are "... about me" at the end of a declarative sentence in a falling intonation. Pitch track and spectrogram generated by Praat (Boersma, 2001).

The longer-term perturbation effects, in contrast, seem to be more directly related to voicing. According to Halle & Stevens (1971), in the production of an obstruent consonant the tension of the vocal folds is adjusted to alter phonation threshold of the glottis (Titze, 1994). This helps to stop voicing in the case of voiceless consonants, or to sustain voicing in the case of voiced consonants (Hanson & Stevens, 2002). These changes in vocal fold tension affect F_0, raising it for the voiceless consonants, and lowering it for the voiced consonants.

Note that the three kinds of vertical effects are separable only when entire F_0 contours of syllables with obstruent consonants are directly compared to reference syllables with an onset nasal consonant. This comparison is based on the assumption that, when an F_0 trajectory is interrupted by an obstruent, the adjustment of vocal fold tension for controlling F_0 does not stop. Continuous adjustment of F_0 without voicing is possible if F_0 and voicing are relatively independently controlled. F_0 is mainly adjusted through vocal fold tension (Fujisaki, 1983), which is done by rotating the thyroid cartilage at its joints with the cricoid cartilage (Hollien, 1960). This involves the antagonistic contraction of the cricothyroid (CT) and the thyroarytenoid (TA) muscles, supplemented by adjustment of laryngeal height and subglottal pressure, with the contraction of the thyrohyoid, sternohyoid and omohyoid muscles (Atkinson, 1978; Erickson et al., 1995). Voicing control, on the other hand, is done by the abduction and adduction of the vocal folds, which mainly involves the lateral cricothyroid (LCT) and the inter-cricoid muscles (Zemlin, 1998).

An additional minor effect is that of aspiration. In an aspirated obstruent, a high rate of airflow is needed to generate sufficient aspiration noise. As a result, part of the transglottal pressure is depleted by the time the glottis closes for the vowel. As a result, the amount of F_0 upshift at voice onset is smaller after aspirated consonants than after unaspirated consonants, as can be seen in Figure 3a for Mandarin (Xu & Xu, 2003). Note, however, this effect occurs very briefly and only at the voice onset. Thus it is not the case that aspiration has a general F_0 lowering effect.

2.3. Pre-low raising

As shown in Figure 2, although the largest effects of the tonal alternation in the second syllable are on the following tone, some smaller effects can be also seen in the preceding syllable. The direction of these anticipatory effects, however, is mostly dissimilatory. That is, the F_0 of the first syllable is the highest when the tone of the second syllable is Low or Rising. This phenomenon, also known as anticipatory raising, anticipatory dissimilation or H raising, has also been found for Yoruba (Connel & Ladd 1990; Laniran & Clements, 2003), Thai (Gandour, Potisuk & Dechongkit, 1994), Igbo (Liberman et al., 1993), Fuzhou (Li, 2015) and many African languages (Schuh, 1978). There is even recent evidence that the height difference between accented and unaccented morae in Japanese could be due to anticipatory raising (Lee, Xu & Prom-on, 2017). The mechanism of this phenomenon is still unclear, but it is likely related to the production of low pitch, which involves a large downward movement of the larynx achieved by the contraction of the external laryngeal muscles, particularly the sternohyoids (Atkinson, 1978; Erickson et al., 1995; Honda, Hirai& Shimada, 1999). The laryngeal lowering shortens the vocal folds by pulling the larynx across a frontward spinal curvature at the neck, which tilts the cricoid cartilage forward, hence moving the arytenoids closer to the front tip of the thyroid cartilage. This shortening is partially passive, and so is in need of a counter contraction of an antagonist muscle. The only muscle that directly lengthen the vocal folds is the cricothyroids (CT) (Zemlin, 1998), whose preparatory contraction may thus raise pitch before it drops into the lower range. There are also other possible mechanisms that have been suggested (Gandour et al., 1994; Lee et al., 2017).

The perceptual effect of pre-low raising has not yet been extensively studied. But there is evidence that speakers actually try to reduce its magnitude during normal production, because when under pressure of heavy cognitive load, the amount of pre-low raising is increased rather than decreased (Franich, 2015).

2.4. Post-low bouncing

Another mechanism closely related to the production of low pitch is *post-low bouncing* (Chen & Xu, 2006; Prom-on, Liu & Xu, 2012). This is the phenomenon that, after a tone with a very low pitch, F_0 sometimes *bounces back* before returning to a normal level. The bounce can take a few syllables to complete and is the most likely to occur when the post-low syllable is weak, e.g., when it has the neutral tone (Chen & Xu 2006), as shown in Figure 5. When the following syllable is not weak, the effect is observable only when the low-pitched syllable is under focus (Xu, 1999). Such a bouncing could be due to a temporary loss of antagonistic muscle balance when the cricoid cartilage returns up across the spinal curvature (Honda et al., 1999), resulting in an abrupt increase of vocal fold tension. This post-low bouncing mechanism has been simulated by increasing the amount of pitch acceleration (second derivative of F_0) in the initial state of the tone that immediately follows the low tone (Prom-on et al., 2012).

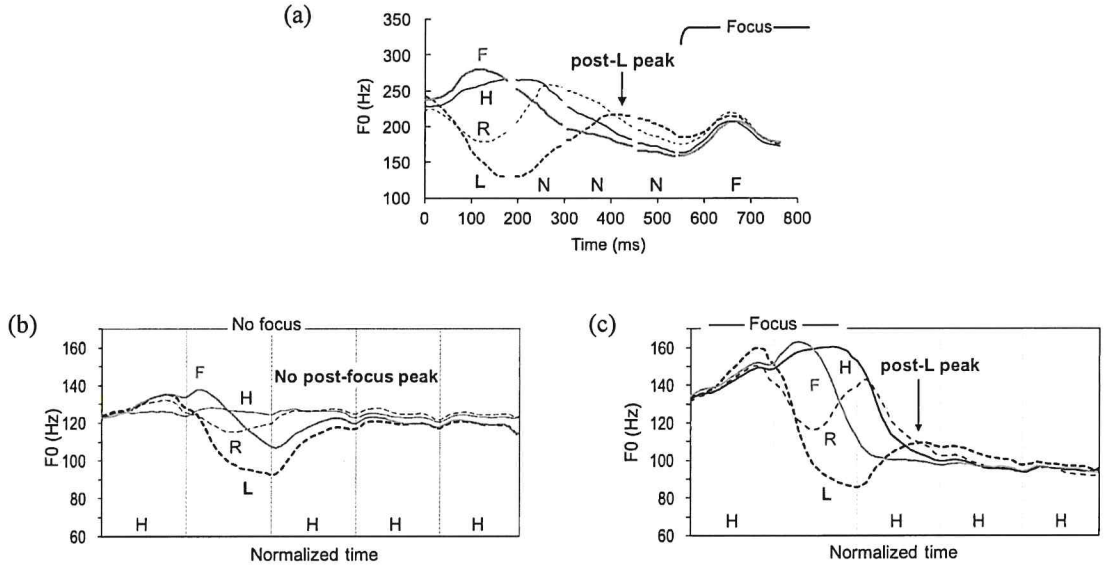


Figure 5. a) Post-low bouncing in Mandarin neutral tone following the Low tone. b) No post-low bouncing in Mandarin full tones; c) Post-low bouncing in Mandarin full tones when the preceding Low tone is focused. Data from Chen & Xu (2006) and Xu (1999).

Post-low bouncing may explain why the Mandarin neutral tone is high pitched after the Low tone but relatively low-pitched after the other full tones (Chao, 1968). The bouncing mechanism could be behind some other tonal and intonational phenomena as well. For example, the distance between a very low-pitched tone in American English and the following F_0 peak remains largely constant (Pierrehumbert, 1980). This parallels the a similar (though a bit longer) constant distance between the Low tone and the F_0 peak in the following neutral tone sequence (Chen & Xu, 2006). Further research in this direction should be able to assess the generality of this phenomenon.

2.5. Total pitch range

According to (Zemlin, 1998), a speaker's conversational pitch range spans about 2 octaves (24 st). Data from a more recent study show that the mean non-singing pitch range of American English speakers, even when not including the falsetto register (Svec, Schutte & Miller, 1999), is about 3 octaves (35.4 st) (Honorof & Whalen, 2005)[1]. In Mandarin, F_0 variations due to the four lexical tones span only about one octave at any particular sentence position (Xu, 1999). Also, the tonal pitch range is in the lower portion of the total pitch range. In (Xu, 1999), the highest mean pitch in sentences with no focus is about 140 Hz for males and 300 Hz for females. With focus it is about 164 and 342 Hz for males and females, respectively. These are still much lower than the highest mean F_0 in (Honorof & Whalen, 2005): 340 Hz for males and 500 Hz for females.

The lower pitch limit, in contrast, is frequently reached in normal speech, as is evident from the diverse strategies speakers employ at the low pitch extreme. For example, for the Mandarin Low tone, as pitch continues to drop, some speakers engage a creaky voice, some use a voiceless whisper, and some even cease phonation altogether, replacing it with a glottal stop in the middle of the syllable (Zheng, 2006). The low-pitch-related non-modal phonations occur not only in the Low tone, but also

[1] Courtesy of Douglas N. Honorof for the numerical data.

in the lowest portion of the Rising tone, and they also become more frequent when the low portions of these tones become even lower under focus (Zheng, 2006). The creaky voice occurs in other languages as well when pitch becomes very low (Redi & Shattuck-Hufnagel, 2001). It thus seems that, in most of these cases, approaching the lower pitch limit leads to the changes in voice quality (Titze, 1994) rather the latter being the primary property of the low lexical tone or boundary tone. In some languages, however, low-pitch related phonation could become the primary feature of some tones. Vietnamese tones, for example, has been argued to show more stable voice quality contrasts (modal, breathy and creaky) than F_0 patterns (Pham, 2003).

The finding that there are about 3 octaves available to the speaker (Honorof & Whalen, 2005) may put to rest a dated, yet still somewhat influential, belief that if a language is tonal, there is not much room left for intonation (Pike, 1948). But pitch range alone is not enough to explain the simultaneous encoding of so many suprasegmental functions. As will be discussed next, virtually all aspects of the dynamic coding mechanism are employed in one way or another by various functions, and even certain aspects of segmental articulation are involved.

3. Parallel Encoding of multiple prosodic functions

While it is no secret that a rich set of meanings are conveyed suprasegmentally, it is often not fully obvious how many different functions are crowded into prosody, as they are typically investigated in separate studies. This section will provide an overview of the full scope of the multiplicity of the functions beyond consonants and vowels, and demonstrate how they are each effectively encoded via non-segmental means. The functions will be divided into three broad categories based on the kind of meanings they convey: *lexical contrast*, *post-lexical linguistic functions*, and *paralinguistic prosody*. As will be seen, most of the functions are multidimensional, but different functions may involve different combinations of the dimensions, and some may heavily rely on only one specific dimension.

3.1. Lexical contrast

This entry point into the prosodic domain of speech is actually overlapped with the segmental domain. Here the suprasegmental cues serve the same function as segments: to distinguish different words, or in some cases, to distinguish different morphosyntactic categories. The best known of these cases are tone and pitch accent, which are used in roughly half of the world's languages (Yip, 2002). Much less emphasized is that word stress is also a lexical function that serves to make some lexical distinctions. What makes stress different from tone and pitch accent is that it is heavily multidimensional, involving vowel color, pitch, duration and intensity all as important cues. The multiplicity of cues is best demonstrated by the classic study of Fry (1958), which examined the perception of near minimal noun-verb pairs in English that are distinguished mainly by stress, e.g., *subject*, *contract*. Interestingly, of the three suprasegmental cues examined—pitch, duration and intensity, pitch is by far the most robust. Only 5 Hz, which is about 0.87 semitones higher than the reference frequency of 97 Hz used in the study, is sufficient for listeners to *unambiguously* hear a word as either a noun (high-low pitch pattern) or a verb (low-high pitch pattern). Nevertheless, pitch is only one of the main cues for stress, beside duration and intensity as found by Fry (1958), and vowel color, as is well known, and the exact repertoire of cues is language dependent (de Jong & Zawaydeh, 2002; Ortega-Llebaria & Prieto, 2011; van Heuven & de Jonge, 2011).

The multiplicity of cues for lexical contrast will be further examined in the following subsections. In particular, given the basic articulatory mechanism of tone production outlined in Section 2.1, there are a number of degrees of freedom that are used by languages to make lexical contrasts. These include target strength, target shape, target timing, target shift, etc.

3.1.1. Target strength as lexical cue

In many cases, a phonological unit does not seem to have clearly specified phonetic properties, because its surface form depends heavily on the phonetic properties of adjacent units. A case in point is the neutral tone in Mandarin. As can be seen in Figure 6b, the F_0 of the neutral tone has very different contours when preceded by different tones. Traditionally, the neutral tone is therefore described as *toneless* (Chao, 1968) or tonally *unspecified* (Chen, 2000). But extensive F_0 variability is also seen at the beginning of the Falling tone in the second syllable after different tones in Figure 6a, which has been established earlier as due to inertia. By the end of the second syllable in Figure 6a (solid arrow), the diverse F_0 contours have all converged to a shape that resembles the phonological description of the Falling tone. A convergence trend can be also seen in the neutral tone on the second syllable in Figure 6b (dashed arrow), however. That is, well before the end of the syllable, all the F_0 contours have started to turn toward a common mid-level value, although full convergence is not achieved by the end of the second syllable. Figure 6c and 6d further show, however, that the convergence continues (dashed arrows) if the third and fourth syllables also have the neutral tone. The only exception is due to the post-low bouncing effect after the L tone as already discussed in 2.4. But even in that case, when the effect is over, F_0 again turns toward the common mid-level height. By the end of the third neutral tone (rightmost arrow, Figure 6d), however, substantial differences related to the tone of the first syllable still remain. Thus, a key difference between the full tones and the neutral tone is that it takes much longer for the latter to reach a target.

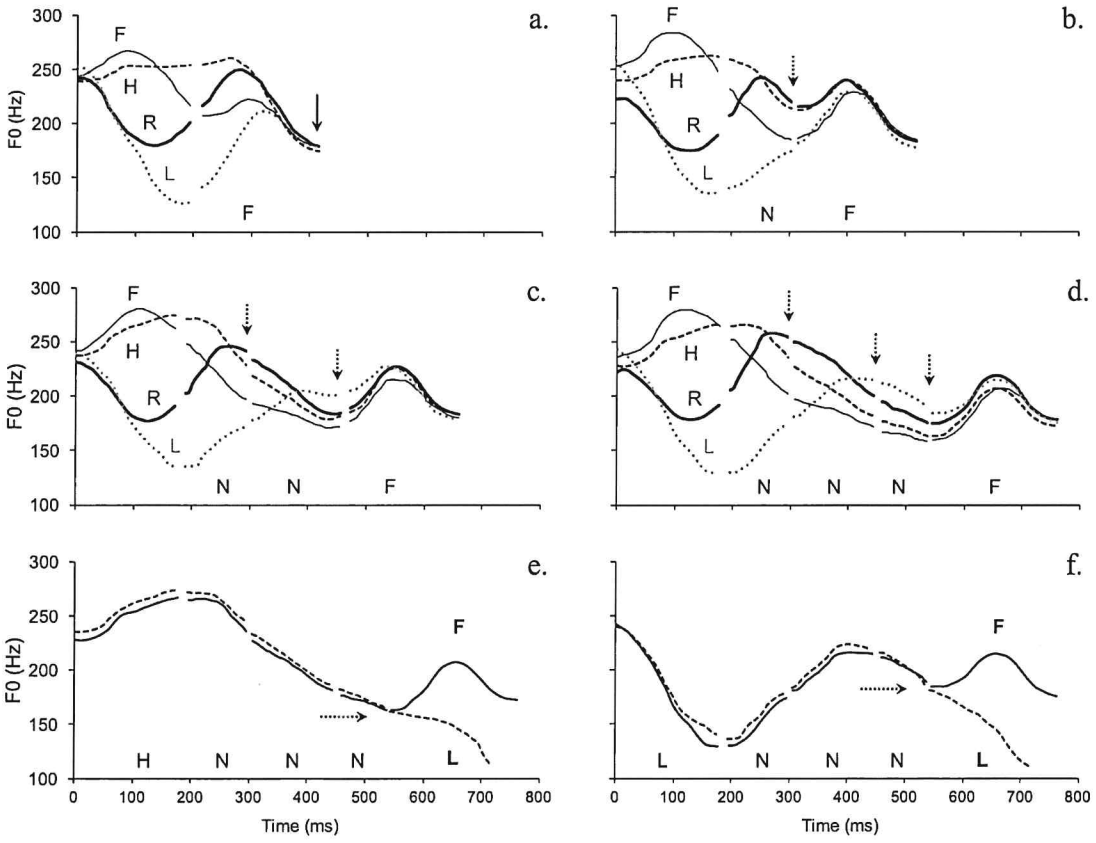


Figure 6. Mean F_0 contours of Mandarin sentences containing 0-3 neutral tone (N) syllables. In (a)-(d), the tone of the first syllable alternates across H, R, L and F. In (a) the tone of the second syllable is Falling. In (b)-(c), there are 1-3 neutral tone syllables after the first syllable. In (e) and (f) the tone sequences are the same except in the last syllable, which alternate between Falling and Low. The horizontal arrows point to a common height reached by the final neutral tone regardless of whether the following tone is Falling or Low. The gaps in the contours represent syllable boundaries. Data from Chen and Xu (2006).

These patterns suggest that the neutral tone also has a pitch target, which is at the mid-level of the tonal pitch range (Chen & Xu, 2006). This can be seen in Figures 6e and 6f, where the third neutral tone reaches a pitch level (indicated by the horizontal arrows) that is halfway between the highest point in the Falling and lowest point in the Low tone in the following syllable. But the slow rate of approaching this target suggests that the tone is produced with a much weaker articulatory force than in the case of a full tone. Based on the estimate of a computational simulation, the strength of the neutral tone is less than half of the full tones (Xu & Prom-on, 2014). It has also been argued that the tone sandhi behavior of Shanghai Chinese involves a weak articulatory strength (Chen, 2008), because similar slow convergence as in Mandarin neutral tone is observed.

Weak articulatory strength may not be limited to lexical tones only. Evidence of weak articulatory strength is also found for unstressed syllables in English (Xu & Xu, 2005) in that although the F_0 of an unstressed syllable is extensively affected by the F_0 of the preceding syllable, the influence fades away over the course of the weak syllable. Yet the rate of reduction of the carryover influence is much slower than in a stressed syllable. The same study also found evidence that the pitch target of unstressed syllables is probably mid, just like in the Mandarin neutral tone. Weak articulatory strength may be used not only in the pitch production, but also in the articulation of segments. Target-like behavior is already demonstrated (Browman & Goldstein, 1992), but explorations as in Chen & Xu (2006) have never been done. This is partly due to an intrinsic difficulty in the segmental domain, because it is hard to have a sequence of unstressed vowels without intervening consonants. The frequent alternation of open and closed phase of the oral cavity would interrupt any continuous target approximation movement toward a neutral target. So, a clever design is needed to systematically study weak segmental targets.

Finally, despite the evidence for weak articulatory strength as a lexical cue as just discussed, it is not highly likely that greater than normal strength is used by any function as a major cue. This is because, as seen in 2.1, maximum speed of pitch change is already often applied in the full tones, which means that maximum articulatory strength is already used. On the other hand, stressed syllables as well as prosodic focus to be discussed in 3.2.1, where a greater strength might reasonably be involved, are known to have increased duration. The duration increase would already provide sufficient time for the targets to be reached. If maximum articulatory strength is still applied, the target might be overshot, which would be undesirable. Given that so far there is no empirical evidence either way, however, further research is needed.

3.1.2. Target shape, contour tone and alignment as cue

The basic dynamics of tone articulation outlined in 2.1 suggests that a target can be as simple as a linear function with an optional slope. This does not rule out the possibility that the target can be more complex. It is also possible that the target can be simply a turning point and its temporal alignment (Dilly & Brown, 2007; Remijsen & Ayoker, 2014). To determine whether a more complex or a simpler target is needed, however, the basic articulatory dynamics still needs to be taken into consideration. For example, due to inertia, the surface form of a unit cannot resemble its underlying

form, because a significant portion of it, especially the early section, is likely to be still far away from the underlying form. This is particularly relevant for interpreting the presence and shape of a sizable movement in the early portion of a contour, and the presence and alignment of a turning point. In Figure 7, for example, the three Mandarin tones all show a sizable rising movement as well as a clear turning point when they are surrounded by two Low tones. But a reexamination of Figure 2 tells us that the rising movement is optional for the High tone and the turning point is optional for both the High and Rising tones. Thus it is unlikely that either the initial rise or the turning points is the main cue for differentiating these three tones.

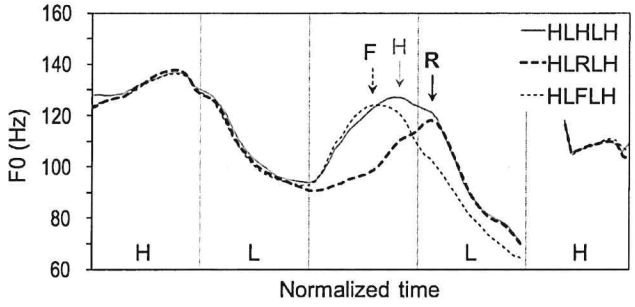


Figure 7. Time-normalized mean F_0 contours of five-tone sequences in Mandarin. The vertical arrows point to the F_0 peaks. All curves averages of 24 repetitions by 4 male speakers. Data from Xu (1999).

It is possible, however, based on the observation of Figure 2, that the Rising and the Falling tones are each composed of two underlying pitch registers: LH and HL, respectively (Duanmu, 1994; Odden, 1995; Yip, 1989), instead of sloped linear functions as suggested by the arrows in the figure. In fact, composite representation of dynamic tones is a widely-accepted norm in tonal phonology (Goldsmith, 1990; Hyman, 2007). Here again, the basic tonal dynamics could offer a way of testing the two accounts. Given that the production of a tone is to approximate its target(s) by overcoming inertia, if there are two consecutive targets, as the time available for a tone increases due to syllable lengthening, the separate approximations of the consecutive targets would become increasingly apparent, as illustrated by the dashed line in Figure 8a. If there is only a unitary sloping target, the slope itself will be increasingly better achieved. When speech rate is directly controlled, the latter is found to be the case for Mandarin Rising tone (Xu, 1998, 2001). That is, as the syllable became longer, it is the rising or falling slope of the dynamic tones that remains the most constant, and there is no sign of separate approximations of consecutive targets. Similar observations have been made for the dynamic tones in Cantonese (Wong, 2006). Interestingly, in both cases, the dynamic portion of the tone is delayed to the end of the syllable, as illustrated by the solid line in Figure 8a.

The delayed realization of a sloping tone could be part of the basic dynamics of target approximation, because other alternatives may not really be viable, as illustrated in Figure 8b. There, the dashed line represents a simple slowing down strategy, which would make the slope increasingly shallower as the syllable lengthens. The dotted line represents a strategy that faithfully realizes the target slope (represented by the arrow), but to do so as soon as it is articulatorily possible. This would result in an increasingly longer final plateau as the syllable continues to lengthen. Neither scenario has been observed so far. Instead, in Thai—a language with vowel length contrast, for example, the most characteristic portion of a tone is realized after a quasi-plateau when the vowel is long (Zsiga & Nitisaroj, 2007). This finding echoes what happens in Mandarin and Cantonese when a syllable with a dynamic tone is lengthened due to a change of speech rate (Xu, 1998; Wong, 2006), as just mentioned. The delayed realization a dynamic target may not be limited to tones. As found by Gay

(1968), the final formant slope of a diphthong is also what remain the most constant as speech rate is reduced.

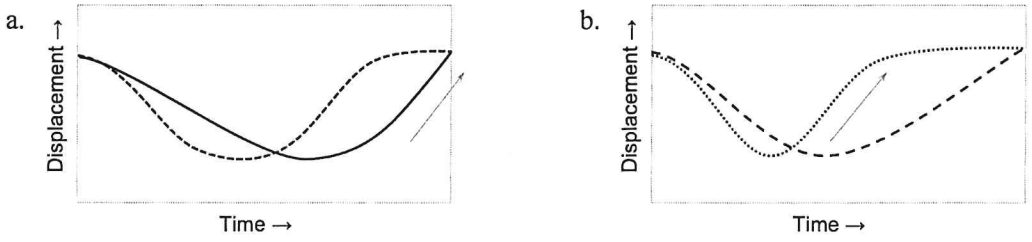


Figure 8. Hypothetical strategies for a contour tone. a) The dashed line results from a sequential approximation of two successive targets; the solid line results from a delayed approximation of a sloping target. b) The dashed line results from a simple slowing down of articulation rate; the dotted line results from a faithful realization of a sloping target at the earliest time allowed by the maximum speed of articulatory movement.

It is also possible that the delay of a dynamic target can itself be used as a distinctive cue. In Shilluk, for example, two falling tones differ from each other mainly in terms of the alignment of the sharpest fall in the syllable (Remijsen & Ayoker, 2014). This is consistent with the idea that speakers are able to delay the onset of a dynamic tone even if there is no alignment contrast (Figure 8a), which is a new direction that can be further explored.

The basic articulatory dynamics of tone production also does not rule out the possibility that there may be cases of genuine consecutive targets within a syllable. An example is seen in Swedish (Figure 9), where the first part of accent II (HL) seem to show a fall followed by a low target (Botinis, Ambrazaitis & Frid, 2014). The falling target is evident in that F_0 first goes up at the beginning of the syllable to reach a high peak, just like in a Falling tone (Figure 7), before dropping sharply. The subsequent low target is evident in that F_0 has dropped to the bottom of the pitch range and formed a low plateau, which is unlike the Falling tones in Mandarin and Thai that drop only to the middle of the pitch range and end with a high falling velocity at the syllable offset (Gandour, Potisuk & Dechongkit, 1994; Xu, 1997), as can be seen in Figures 2 and 7. Again, this is yet another new direction for further research.

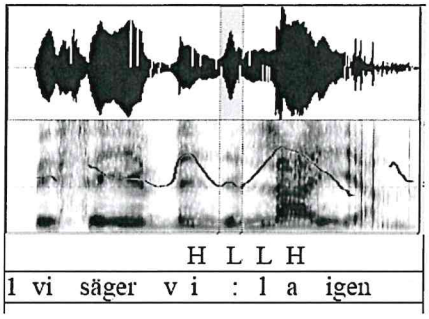


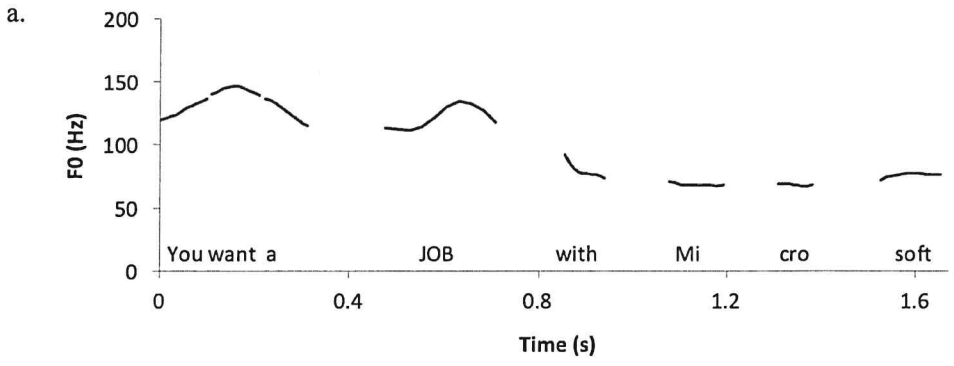
Figure 9. Swedish Accent II, where a Fall-Low sequence can occur within the same vowel [i:] in the word "vi:la". Courtesy of Antonis Botinis.

3.1.3. Target reassignment

The basic articulatory dynamics can also help to identify cases where tonal variations are *unlikely* to be due to inertia. A clear example is the Low-tone sandhi phenomenon in Mandarin (Chao, 1968), whereby the first Low in a Low-Low sequence is changed into a form that sounds indistinguishable from the Rising tone (Peng, 2000; Wang & Li, 1967). Despite many efforts, it has been difficult to identify a mechanism that can unequivocally derive this phenomenon. Detailed acoustic analyses show that the sandhied F_0 contour approaches a rising target even from the syllable onset, just as in an underlying Rising tone (Xu, 1997). Thus there seems to be a genuine *reassignment* of the underlying pitch target, although the exact reason for the reassignment is still unclear.[2]

The arbitrariness of the tone sandhi rules has been well recorded (Chen, 2000). Recent research is able to establish further that some of the arbitrary rules are still relatively transparent in that they are applied by speakers whenever the phonetic environment satisfies. In contrast, there are also sandhi rules that are *opaque*, as they are applied only to words that speakers are familiar with. The latter mechanism has been termed lexical or allomorph listing (Tsay & Myers, 1996; Zhang, 2014; Yip, 2004). It has been further demonstrated that different strategies may be involved in learning these types of tone sandhi rules. Currently this is a very active area of research (Zhang, 2014), which is likely to advance our knowledge significantly about the nature of tone sandhi.

What has been much less discussed is that tone-sandhi-like rules may apply in non-tone languages as well. In Figure 10a, for example, the monosyllabic word *job* shows a rise-fall contour. A similar rise-fall contour can be also seen in *Bloomingdales* in Figure 10b, but the shape is distributed across the whole word. This similarity may indicate a functional equivalence of the two contours (Arvaniti & Ladd, 2009; Ladd, 2008; Pierrehumbert, 1980). But the equivalence appears to be achieved by assigning a high-level or rising target to the stressed syllable, and a mid- or low-level target to the unstressed syllables. Thus the fall in the monosyllabic word is split into a sequence of targets that are assigned to the individual syllables of a multisyllabic word. Such target *reassignment* is therefore somewhat similar to the target reassignment in tone sandhi. Evidence for similar target reassignment is found also for the interrogative intonation (Liu et al., 2013), as will be further discussed in connection with focus and sentence modality in 3.2.1 and 3.2.2.



[2] There is some evidence that this is a historical change that started centuries ago and slowly spread across many northern dialects (Mei, 1977). This makes the true mechanism of the change hard to unearth, as it is difficult to know the exact pitch value of the Low tone at that time.

b.

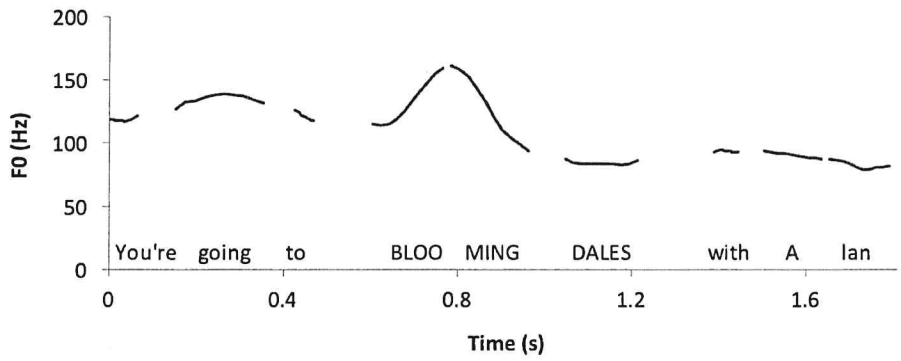


Figure 10. Mean F_0 contours of a) You want a JOB with Microsoft. and b) You're going to BLOOMINGDALES with Alan, averaged across 8 repetitions by a male American English speaker. Capitalized words are in focus. Data from Liu et al. (2013).

The reassignment of the targets is likely necessitated by the nature of speech articulation. That is, it is impossible to separately articulate a sequence of pitchless syllables and a sequence of segmentless/syllableless F_0 contours and then somehow combine them later on, as assumed by some computational models of intonation (Fujisaki, 1983; van Santen et al., 2005). Instead, regardless of what a global contour may look like, each syllable in a multisyllabic word or phrase has to have its own underlying pitch target while it is being articulated, just as it has to have its own consonantal and vocalic targets. Besides, the syllable-specific targets need to be specified in terms of stress-related articulatory strength, as discussed in 3.1.1, which cannot be part of the specifications of a functional equivalent global contour. Target reassignment is therefore probably much more widespread across languages than have been recognized. Further research may consider it as a possible factor when trying to account for various tonal and intonational phenomena.

3.2. Post-lexical Linguistic functions

Prosody is also heavily involved in conveying meanings that are independent of lexical distinctions. As a result, the temporal domain of these functions is not limited to syllables or even words. Yet, for reasons just discussed in 3.1.3, even the prosody of larger temporal domains has to be ultimately realized via individual syllables. But here we are concerned mainly with the more global prosodic patterns, and the basic articulatory dynamics will be mentioned only in cases where it plays a critical role.

A main issue in regard to post-lexical prosody is that of function versus form. There is a strong form-oriented tradition which assumes that the understanding of prosody should start from a thorough description of the observable prosodic forms. The British school of intonation, for example, formulates a general scheme in which the intonation of any utterance consists of an obligatory nucleus and an optional head, prehead and tail (O'Connor & Arnold, 1961; Palmer, 1922). Likewise, intonational phonology, also known as the Autosegmental-Metrical phonology theory of intonation (AM theory), formulates a scheme in which prosody is guided by a finite state grammar (Gussenhoven, 2004; Ladd, 1996; Pierrehumbert, 1980). In both cases, meanings are associated to the prosodic forms only *after the forms have been established* in terms of both their prosodic properties and rules that guide their alternation. As has become increasingly clear from the discussion so far, there seem to be many degrees of separation between the meaning-carrying functions and

their associated prosodic form. First, due to inertia, an underlying target and its surface realization often look only faintly like each other. Second, a function-specific form may often assign different underlying targets depending on various factors, as discussed in 3.1.3. Finally, there is also a third way of function-form separation, namely, the simultaneous encoding of multiple functions. Consequently, no temporal location of an utterance can be said to correspond to any single prosodic function. It is therefore ineffective to first categorize directly observable forms and then try to link each of them to a specific meaning-bearing function.

The discussion of post-lexical linguistic prosody in this section will follow, instead, a function-oriented approach. Each of the subsections will address whether and how a particular hypothetical function is encoded with specific prosodic cues. And the evidence cited will also be mainly from studies that have systematically controlled the prosodic functions being examined.

3.2.1. Focal contrast

Despite being one of the most widely discussed prosodic functions, focus still is in want of a precise definition. Earlier theories tend to associate focus with new information (Chaffe, 1976; Halliday, 1967), but exceptions as in the following examples make it untenable.

(3) A: Did John or Mary have an operation? B: *Mary* had an operation.

(4) A: Why did you miss the party? B: My *mother* got sick.

In (3) "Mary" is likely to be focused although it is already mentioned in the preceding question and thus is given. In (4) "got sick" is unlikely to be focused although it has not been mentioned before, as pointed out by Terken and Hirschberg (1994). Furthermore, new information is virtually everywhere if an utterance is to convey any meaning. A more sophisticated definition has been that focus is for indicating that there is an alternative relevant to the interpretation of a given linguistic expression (Krifka, 2008; Rooth, 1992). But an alternative-based definition would still over-predict focus because it is not the case that whenever there is an alternative interpretation in an utterance focus has to occur. A more restricted definition would be based on the need to emphasize something. That is, when focus occurs, the speaker has felt a need to emphasize to the listener what is particularly important in the utterance, and is worthy of special attention. Evidence for this account is shown by recent studies using fMRI, ERP and eye tracking technology (Chen, Li & Yang, 2012; Chen, Wang & Yang, 2014; Kristensen et al., 2013). The findings of those studies suggest that focus is associated with attention allocation during discourse comprehension, while newness is associated only with memory retrieval.

For the purpose of empirical research, what is important is to have methods that can reliably elicit focus. Experimental studies have shown that mini dialogues are a highly reliable way of inducing focus. In a mini dialogue, the leading sentence, or phrase, can be either a wh-question or a statement containing a piece of information to be negated in the target utterance. The former case is often referred to as information focus while the latter as contrastive focus.

One of the most classical examples of well controlled experiments on focus is a series of studies by Cooper, Eady and colleagues, which used the mini dialogue paradigm (Cooper, Eady & Mueller, 1985; Eady & Cooper, 1986; Eady et al., 1986). These studies established that focus in American English is prosodically realized mainly by a) consistently decreased peak F_0 of post-focus words, b) consistently increased duration of the focused word, and c) occasionally increased peak F_0 of the focused word. These are confirmed by later studies that examine continuous F_0 contours in greater detail (Xu & Xu, 2005; Liu et al., 2013). Some examples are shown in Figure 11a-b. Also using the

IPR2026-00273
Krisp Technologies EX1025 Page 16

mini dialogue paradigm, Jin (1996) and Xu (1999) found similar post-focus lowering of F_0 and on-focus increase of duration in Mandarin Chinese. But for the focused word in Mandarin, the effect is an *expansion* of the pitch range, consisting of increased F_0 maximum and decreased F_0 minimum, as can be seen in Figure 11c. Later studies further found intensity variations similar to those of F_0: on-focus increase and post-focus reduction (Chen, Wang & Xu, 2009; Lee et al., 2015). Thus for both English and Mandarin, the most consistent correlates of focus are post-focus compression (PFC) of F_0, and on-focus increase of duration.

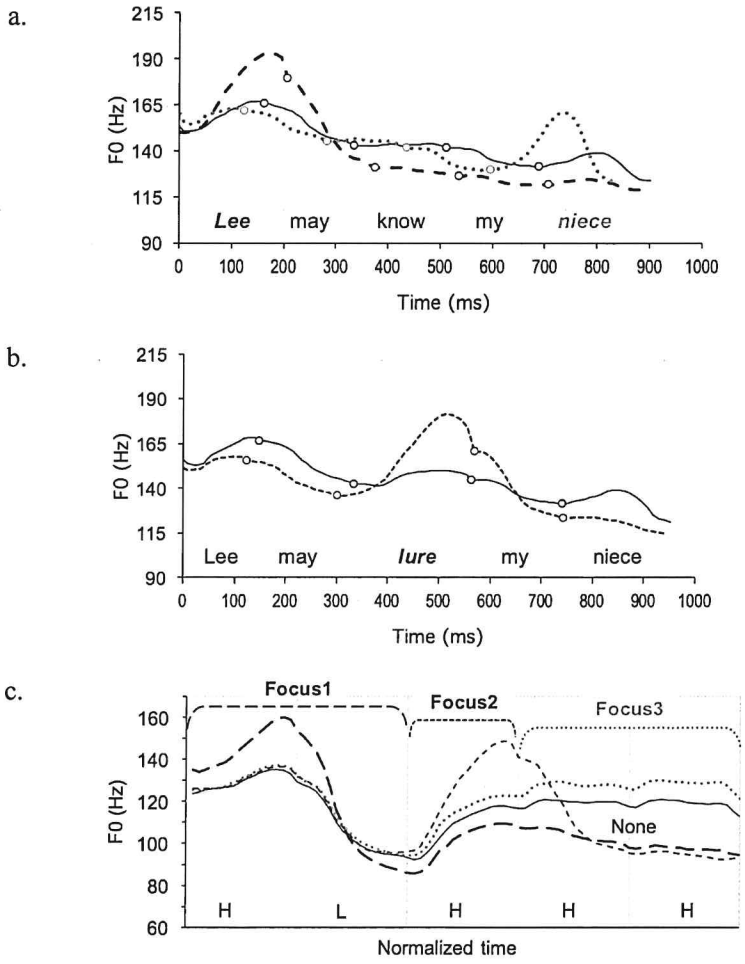


Figure 11. Focus prosody in English (a, b) and Mandarin (c). Data from Xu and Xu (2005) and Xu (1999).

Similar prosodic patterns, especially PFC of F_0, have been found in many other languages, including German (Féry & Kügler, 2008), Dutch (Rump & Collier, 1996), Swedish (Bruce, 1982), Japanese (Ishihara, 2002), Korean (Lee & Xu, 2010), Turkish, (Ipek, 2011); Uygur (Wang et al., 2011), French (Dohen & Lœvenbruck, 2004), Arabic (Chahal, 2003), Hindi (Patil et al., 2008), Persian (Taheri-Ardali & Xu, 2008) and Finnish (Mixdorff, 2004). These languages all belong to major Euro-Asian language families: Indo-European, Uralic, Afro-Asiatic and Altaic. The only exceptions are Mandarin (Xu, 1999) and Tibetan (Wang, Wang & Xu, 2012), which belong to the

Sino-Tibetan family. At least for some of these languages PFC is shown to be critical for focus perception: Finnish (Vainio et al, 2003), Dutch (Rump & Collier, 1996), English (Prom-on, Xu & Thipakorn, 2009), Japanese (Ishihara, 2011; Sugahara, 2005), and Mandarin (Liu & Xu, 2005, Chen et al., 2009).

But PFC is absent in many other languages, including Southern Min (Chen, Wang & Xu, 2009), Cantonese (Wu & Xu 2010), Deang, Wa and Yi (Wang, et al., 2011), Yucatec Maya (Kügler & Skopeteas, 2007), Chichewa, Chitumbuka, Durban Zulu, Hausa, Buli and Northern Sotho (Zerbian, Genzel & Kügler, 2010) and Wolof (Rialland & Robert, 2001). Some of these languages even show on-focus increase of pitch range, duration or intensity (Chen et al., 2009; Wu & Xu, 2010). Thus there seems to be a dichotomy in terms of whether a language uses PFC to mark focus, and the division seems to be partially related to language families: The non-PFC languages do not belong to any of the families of the known PFC languages: Indo-European, Uralic and Altaic. The exceptions are the Sino-Tibetan and Afro-Asiatic families, which host both PFC (Mandarin, Arabic) and non-PFC languages (Cantonese, Taiwanese, Hausa). These distribution patterns have led to the hypothesis that languages with PFC are all related in that they are common descendants of a proto-language (Xu, 2011).

Further motivating this *inheritance hypothesis* are two key findings. First, the distribution of PFC is not related to any language specific features, such as tone or lexical stress (Chen et al., 2009). This means that PFC could not have emerged in a language due to these features. Second, PFC is hard to spread across languages, since a) it is not transferred from a PFC language to a non-PFC language even when both languages are spoken natively by bilinguals (Wu & Chung, 2011), or near natively (Chen, Xu & Guion-Anderson, 2014), b) it is hard to acquire by an L2 learners whose L1 has no PFC (Chen et al., 2014; Swerts & Zerbian, 2010), and c) it is hard to acquire even by an L2 learner whose L1 already has PFC (Chen, 2015). This means that a language, e.g., Mandarin, could not have acquired PFC through contact with a neighboring PFC language, such as Mongolian or Manchurian (whose speakers ruled China for hundreds of years), but it itself has to be a descendent of a PFC language (Xu, 2011).

As for the identity of the common-protolanguage, the distribution of PFC across Indo-European, Uralic, Afro-Asiatic and Altaic languages shows a resemblance to the hypothetical Nostratic macro family, according to which many of the Euro-Asian and Afro-Asiatic languages share a common proto-language that originated from the Fertile Crescent (Bomhard, 2008; Pedersen, 1931). The Nostratic macro family hypothesis is also largely consistent with the Farming/Language Dispersal Hypothesis (Diamond & Bellwood, 2003), according to which many of the major language families of the world dispersed along with the spread of agriculture, and the Fertile Crescent is the origin of the earliest agriculture, based mainly on wheat, barley, millet, and emmer (Diamond, 1998).

Probably the most radical aspect of the *inheritance hypothesis of PFC* is that it splits the Chinese languages into two groups, one consisting of many of the northern dialects with PFC, and the other consisting of many of the southern dialects/languages that lack PFC. Dialects in the northern group are likely descendants of Altaic languages, given the non-tranferability of PFC. Dialects/languages of the southern group, on the other hand, have two possible hereditary roots. Some could be descendants of a protolanguage (or protolanguages) originating from an area of the rice-based agriculture (Diamond, 1998). Others may be close relatives of northern Chinese, but have lost PFC because of language shifts in southern populations that originally spoke various non-Chinese languages. Radical as it is, the inheritance hypothesis has generated some exciting new research that may shed light not only on speech prosody, but also on broader issues like language change, language typology and language evolution.

There are also three more issues about focus that are worth mentioning. Firstly, despite the widely-known division between various types of focus (Gussenhoven, 2007), especially between *information focus*—which is elicited by wh-question, and *contrastive focus*—which is used to correct a wrong piece of information, empirical research has not yet generated sufficient support for a clear distinction between them. Most studies have failed to find significant differences (Hanssen, Peters & Gussenhoven, 2008; Sityaev & House, 2003; Hwang, 2012; Katz & Selkirk, 2011; Kügler & Ganzel, 2014). But a few studies have found some small differences (Baumann et al., 2007; He et al., 2011; Sahkai, Kalvik & Mihkla, 2013). As will be discussed later, there is the possibility that even these small differences may not be due to focus type, but are attributable to paralinguistic functions that interact with focus.

Secondly, probably because of the critical role of PFC, final focus is often less effectively encoded than an earlier focus, as PFC is impossible to apply to the sentence-final location (Botinis, Fourakis & Gawronska, 1999; Liu & Xu, 2005; Rump & Collier, 1996). This finding is especially relevant for the controversial notion of *broad focus* that is used to refer to the prosody of sentences with no narrow focus, which are said to have a default sentence-final nuclear accent (Ladd, 1996, 2008). The weak prosodic cue for final focus may explain why utterances with no focus or neutral focus may have sounded as if there were a final accent, but it does not make them equivalent to final focus. The significant differences between broad and final focus has been clearly demonstrated (Katz & Selkirk, 2011; Xu, 1999; Xu & Xu, 2005). Besides, the idea that an entire sentence is focused when spoken as an answer to a question like *What happened?* seems to conflict with the idea of focus as selective highlighting, whether for the sake of pointing out the existence of alternatives (Krifka, 2008; Rooth, 1992) or directing listener's attention (Chen et al., 2014).

Finally, despite being an independent function in its own right, focus interacts with other functions in ways that may affect either the realization of focus itself or that of the other functions. For its own realization, focus is not effectively encoded in unaccented words in Japanese (Lee & Xu, 2012; Ishihara, 2011), and is less effectively realized on the Low tone in Mandarin than on the other tones (Lee, Wang & Liberman, 2016). For the realization of other functions, in American English, a word-final stressed syllable (including that of a monosyllabic word) changes its target from a non-dynamic high to a dynamic fall when under focus (Xu & Xu, 2005; Liu et al., 2013). In Mandarin, in contrast, focus does not change the underlying targets of tone despite the changes in pitch range (Xu, 1999). As will be discussed in the next section, the interaction of focus with the modality function also involves changes in underlying pitch targets.

3.2.2. Modality

Modality, or sentence type, refers to whether an utterance is produced as a statement or a question. Much research has been done on the prosody of modality, but the general picture is not yet fully clear. On the one hand, there is a widely-recognized cross-linguistic trend that a question often involves a rising intonation as opposed to a falling intonation in a statement (Bolinger, 1978; Greenberg, 1963; Ohala, 1983; Shattuck-Hufnagel & Turk, 1996). On the other hand, question-final pitch rise is often missing either in a language known to have rising question intonation, or absent in an entire language or dialect (Savino, 2012; Siemund 2001; Zerbian, 2010). Here again, systematic analysis of continuous pitch contours may help to identify phonetic details that can make the picture clearer. Figure 12 shows mean F_0 contours of statements and questions consisting of words all with the same tones in each sentence (Liu & Xu, 2005). As can be seen, not all questions show sentence-final rises. Although F_0 rises sharply in both the Rising-tone and Low-tone questions, the Falling-tone question shows a sharp final fall just as in a statement, and the final rise in the High-tone question is not very sharp. What is common across all the tones, however, is the increasing pitch level throughout the sentence. As found in the study, the global difference between statements and questions can be fitted

IPR2026-00273
Krisp Technologies EX1025 Page 19

by a double exponential function, which can be considered as consisting of an *accelerating descent* in statements and an *accelerating ascent* in questions.

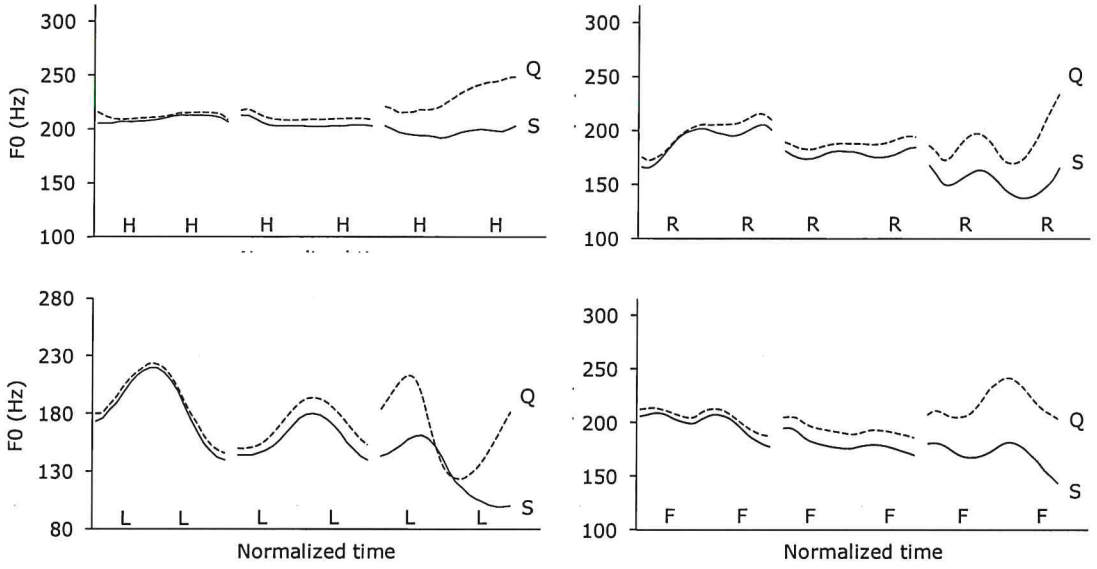
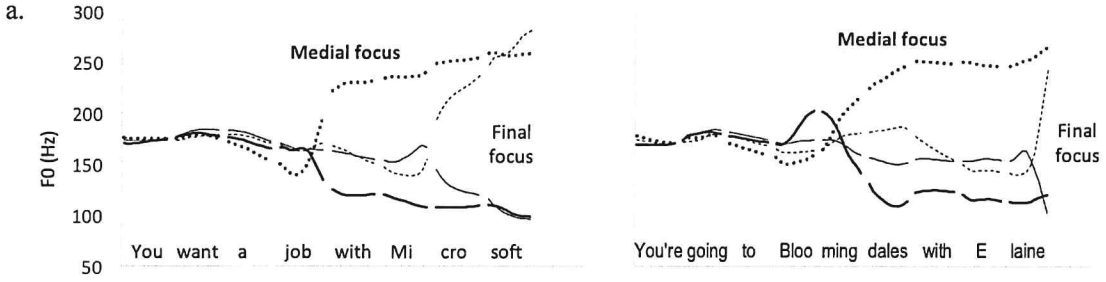


Figure 12 mean F_0 contours averaged across 8 speakers of Mandarin Chinese (5 repetitions by each speaker). Data from Liu & Xu (2005).

So, for Mandarin, the globally accelerating pitch level increase seems to be the *primary* cue for question prosody, which is "imposed" onto the tones of successive syllabic morphemes. Thus the local shape at the end of a sentence mainly reflects the lexical tones rather than being the question intonation proper. There is an exception to this simplistic account, however. That is, the L-tone sentence shows a simple drop of F_0 to the bottom at the end of a statement, but a low-rise contour at the end of a question. These two contour patterns are in fact two of the allophonic variances of the L tone (Chao, 1968). Thus, in addition to the accelerating pitch level, the interrogative function also interacts with the lexical tonal function to reassign the local tonal targets. Such interactions may be quite frequent across languages. In Cantonese, for example, all the tones change to a sharp rise in questions (Ma, Ciocca & Whitehill, 2006), which result in a reduction of tone identification (Fok-Chan, 1974; Ma et al., 2006). Languages, therefore, may differ in terms of whether modality or lexical contrast is given a greater weight in the control of sentence-final F_0 contours.



b.

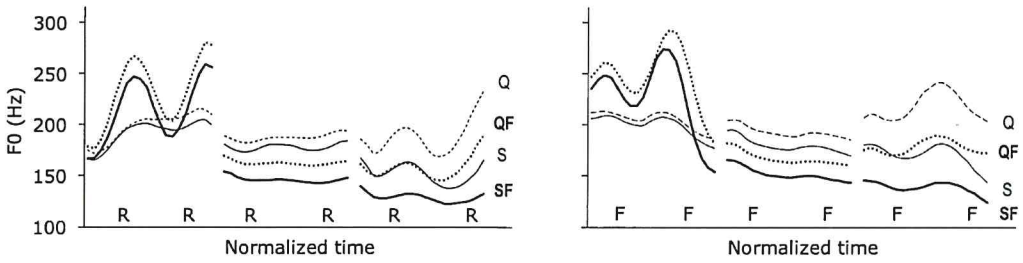


Figure 13. a) Mean F_0 contours averaged across 5 speakers of American English (8 repetitions by each speaker). b) Mean F_0 contours averaged across 8 speakers of Mandarin (5 repetitions by each speaker). Data from Liu et al. (2013).

Even more dramatic interactions of the interrogative function with other functions can be seen in English (Liu et al., 2013). Figure 13a shows mean F_0 contours of two English sentences spoken as statements and questions, with focus in either medial or final positions. Here we can see three-way interactions between focus, modality and lexical stress:

1. The pitch range of words are jointly determined by focus and modality. In a statement, pitch range is compressed and *lowered* after focus, but in a question, pitch range is compressed but *raised* after focus. Thus post-focus compression applies to both modalities, but differ in terms of whether the compressed pitch range is raised or lowered.

2. The local contours of stressed syllables also vary depending on both focus and modality. Before and after focus, they are largely flat, under focus, they are either falling and rising depending on whether the sentence is a statement or questions, respectively.

3. There are also subtle changes in the local contour of stressed syllables that reflect a tendency for their underlying pitch targets to be slightly rising in questions but falling in statements even in pre-focal and post-focal words. This is revealed by both systematic acoustic analysis and modeling simulation in Liu et al. (2013).

These interaction patterns contrast with similar interactions in Mandarin as shown in Figure 13b in a number of ways. First, the post-focal pitch range is lowered in Mandarin even in questions. Second, as discussed earlier, the local tonal shapes in Mandarin do not vary in fundamental ways with changes in modality. But similar to Mandarin as mentioned earlier, the manifestation of modality in English is also global rather than limited only to the sentence final location. The early manifestation is perceptually relevant too. Both adults and 7-10 old children can identify sentence type correctly often as early as from the first word of a sentence (Saindon et al., 2017). Similar early identification of sentence type has been reported for other languages (Face, 2005; Thorsen, 1980; van Heuven & Haan, 2000).

Furthermore, the interaction of modality with pragmatic and paralinguistic functions introduces further complications in the prosody of question intonation. In Figure 14 we can see that the pitch ranges of questions with various pragmatic connotations form a gradience rather than clear-cut categories. Such a gradience cannot be fully accounted for in terms of linguistic prosody only. Also shown in Figure 14 is that sentences with a sentence-final particle have the second highest F_0 peak. This means that question particle and interrogative intonation are not in a trading relation (see also Abe, 1998 for Japanese). Further discussion on this will be resumed in 3.3.

IPR2026-00273
Krisp Technologies EX1025 Page 21

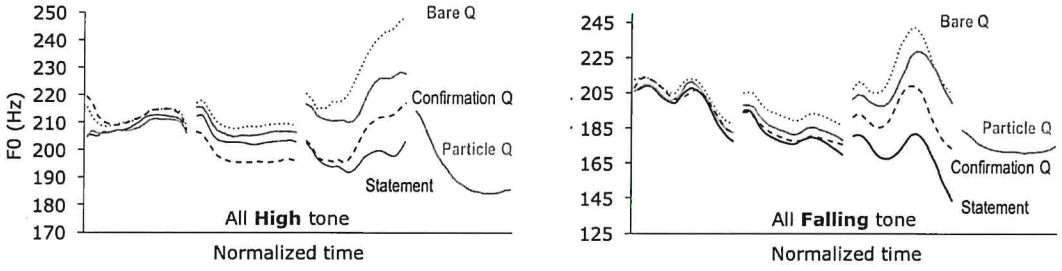


Figure 14 mean F_0 contours averaged across 8 speakers of Mandarin (5 repetitions by each speaker). Data from Liu and Xu (2005).

Finally, the accelerating global increase of pitch range in question intonation may be entirely absent in some languages. In three African language phyla: Niger-Congo, Nilo-Saharan, and Afro-Asiatic, questions are marked by *lax prosody* (Connell, 2017; Rialand, 2009; Salffner, 2010). Instead of sentence-final pitch increase, these languages use properties such as falling pitch contour, a sentence-final low vowel, vowel lengthening, and breathy utterance termination to mark questions. A remote yet intriguing possibility is that the accelerating pitch increase shared by so many of today's languages happened to be in the prosody of the language(s) spoken by people who came out of Africa tens of thousands of years ago, while the languages that remained in Africa always had a variety of question prosody. Thus, just like PFC, which may have split the modern languages from 13,000 years ago, lax prosody may have split languages even earlier, assuming that global prosody is among the most stable features of human speech.

3.2.3. Grouping/boundary marking by duration

The grouping function divides an utterance into chunks, presumably for the ease of production as well as perceptual processing (Cutler, Dahan & van Donselaar, 1997; Schafer et al., 2000). Although many prosodic cues have been reported, there is increasing evidence that it is the temporal domain of speech that provides the strongest and most reliable cues. Temporal cues consist of both the duration of articulated units (vowel, syllable, words, etc.) and the length of silent pauses. There is also evidence of cues in terms of pitch, in the form of pitch reset (Ladd, 1988; Swert, 1997). But they occur only in limited situations, and are often byproducts of communicative functions already discussed. There are also phenomena, such as rhythm metrics and hierarchical prosodic structures, that do not seem to serve any specific functions. But they are likely also byproducts of boundary marking as well as some other meaning-carrying functions (Arvaniti, 2012).

For temporal cues used as boundary marking, evidence comes from rather diverse sources, including domain-final lengthening, polysyllabic shortening and pausing. Domain-final lengthening, or final lengthening for short (also known as preboundary lengthening), refers to the fact that the duration of a syllable, syllable rhyme or the nuclear vowel increases at the end of increasingly larger syntactic/prosodic units (Klatt, 1976). Polysyllabic shortening refers to the finding that a stressed syllable tends to be increasingly shorter as the number syllables in a word or phrase increases, e.g., in *sleep*, *sleepy*, *sleepily*, *sleepiness* (Lehiste, 1972; Klatt, 1976). Given that they seem to be complementary to each other, the two phenomena could be reflections of the same trend of duration variation: the closer to the right edge of a unit, the longer the duration, and vice versa. This is partially evident in Figure 15a for American English (Nakatani, O'Connor & Aston, 1981). From right to left, syllable duration first decreases sharply, and then reaches a low plateau for within word positions. The same trend is also found in Mandarin (Xu & Wang, 2009), but even stronger, as can be seen in

IPR2026-00273
Krisp Technologies EX1025 Page 22

Figure 15b. Here the increase in syllable duration occurs not only on the right edge between word/phrase final syllable and non-final syllable, but also on the left edge, i.e., between word-initial and word medial syllables.

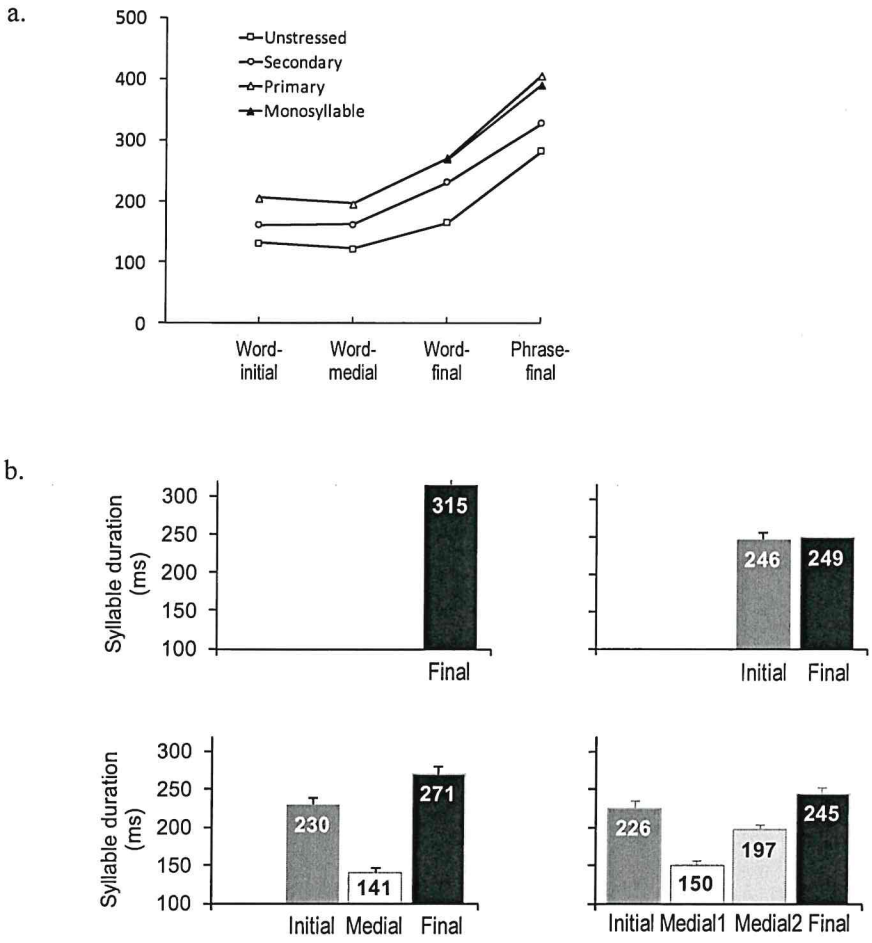


Figure 15. a) Syllable duration as a function of position, stress level and word length in English (redrawn based on data from Figure 4 of Nakatani et al., 1981). b) syllable duration as a function of position and word/phrase length in Mandarin. The numbers on the bars are in ms. Adapted from Xu and Wang (2009).

There is a limit to how much a syllable can be lengthened to signal a boundary, however. The lengthening of a sentence-final syllable is found to be no greater than the lengthening of phrase-final syllables in American English (Klatt, 1975; Umeda, 1975; Wightman et al., 1992; Yang & Wang, 2002). This limit seems to be related to the second temporal cue for grouping, namely, silent pause. Silent pause refers to the full absence of phonation that lasts longer than the duration of any segment-intrinsic silence, e.g., a stop closure. There does not seem to be a limit to how long a silent pause can be. In fact, the length of a silent pause increases quasi-exponentially with the increase of perceived boundary strength, as can be seen in Figure 16 for Mandarin (Yang & Wang, 2002).

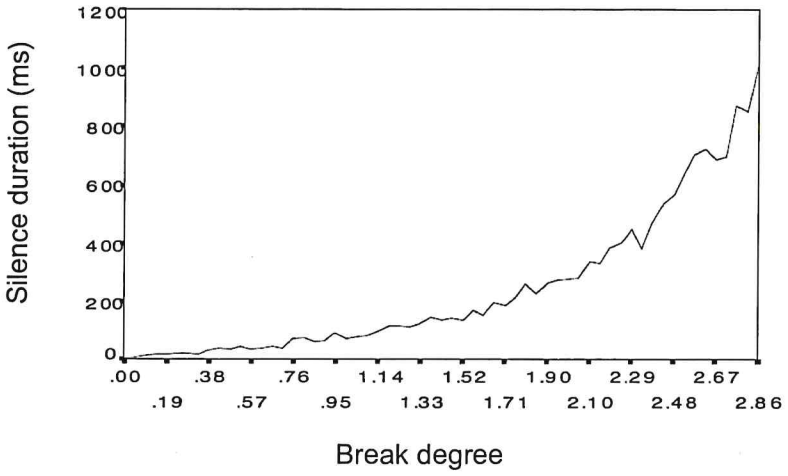


Figure 16. Duration of silence pause as a function of perceived prosodic boundary level (Yang & Wang, 2002). Courtesy of Bei Wang.

Thus there seems to be a division of labor between final lengthening and silent pause in marking boundaries (Petrone et al., 2017). For weaker boundaries, only final lengthening is involved, while for stronger boundaries, e.g., groups larger than phrases, only silent pauses can indicate a further increase in boundary strength. At the phrase level, however, both cues are at work (Jeon & Nolan, 2013). It is therefore possible to combine final lengthening and pausing into a joint measurement of boundary strength, e.g., affinity index (Xu, 2009). Affinity refers to the fact that both domain final syllable and silent pause increase the distance between the onset of final syllable and the onset of the upcoming syllable, and this distance seems to iconically encode the relational distance between the two adjacent domains: the closer the relation, the weaker the boundary between the two, and vice versa.

Also, as it appears in Figure 16, and argued by Wagner (2005) and Xu and Wang (2009), boundary strength is likely gradient rather than categorical. In Wagner (2005), listeners can identify up to seven levels of recursive syntactic relations, each with a boundary strength scaled by duration. In Xu & Wang (2009), even within a phrase, durational patterns reflected whether the phrase-internal grouping was 1+3 or 2+2. Similar sensitivity to boundary cues has also been shown by O'Malley, Kloker and Dara-Abrams (1973) and Fant and Kruckenberg (1996).

The gradiency of temporal marking of boundary strength discussed so far raises questions about the widespread notion of prosodic hierarchy (Beckman, 1996; Ladd, 2008; Selkirk, 1986). According to this notion, speech prosody is a hierarchical structure consisting of a limited number of levels: mora, prosodic word, foot, phrase, intonational phrase, and utterance. This structure not only governs its internal constituents, but also determines their relative prominence (Beckman, 1996; Liberman & Prince, 1977; Selkirk, 1980). The fine gradience of boundary strength makes it hard to find any fixed set of constituents to match a prosodic hierarchy. Furthermore, Figure 15a shows clearly that the effects of lexical stress in English are independent of durational variations due to boundary strength, without any cross interactions (Nakatani et al., 1981). Thus the grouping constituency is unlikely to govern word stress, which is lexically defined. Even further back, we have also seen that focus is functionally independent of lexical stress. It has also been shown that phrasal marking and focus are independently controlled in Japanese (Ishihara, 2011; Kubozono, 2007), and focus condition does not affect pauses in Swedish (Horne, Strangert & Heldner, 1995). Thus there does not seem to be a functional need to use a hierarchical division of grouping to control either lexical stress or focus.

The temporal patterns related to boundary marking as well as lexical stress and focus also call into question the widely-known rhythm class hypothesis (Abercrombie, 1967; Pike, 1945). The hypothesis divides languages into stress-timed (e.g., English, Arabic), syllable-timed (French, Italian, Yoruba) or mora-timed (Japanese), whereby the said timing unit is supposed to show a tendency of equal duration, i.e., equal temporal distance between stresses, syllables, and morae, respectively (Abercrombie, 1967; Port, Dalby & O'Dell, 1987; Steever, 1987). Although such isochrony has been repeatedly demonstrated to be nonexistent (Dauer, 1983; Lehiste, 1977; Nakatani et al., 1981; Warner and Arai, 2001), the hypothesis is revived by the proposal of a number of rhythm metrics that can nevertheless divide or arrange languages along the dimensions defined by these metrics (Dellwo, 2006; Grabe & Low, 2002; Ramus, Nespor & Mehler, 1999). But the term stress-timing or syllable timing should at least be indicative of a weak tendency toward isochronic recurrence of the said unit in the language. Yet even this tendency can be put into question. For example, based on the rhythm metrics, Grabe and Low (2002) and Lin and Wang (2007) have demonstrated that Mandarin firmly belongs to syllable-timed languages. Yet the data shown in Figure 15 indicate that Mandarin syllables are actually more compressible (hence less isochronous) than English, and Mandarin phrase-level units are more adjusted toward isochrony than English. The first trend is seen in the fact that phrase-medial syllables in Mandarin are shorter than the phrase-initial syllables, whereas in English, being word medial does not further shorten the syllable at any specific stress level. The second trend is seen in the fact that, in Mandarin, a syllable in a monosyllabic word becomes much shorter when it is in a disyllabic word, even when it is still word final, whereas there is no significant shortening from monosyllabic word to multi-syllabic word in English. When the two trends are combined, Mandarin shows a greater tendency than English to equalize the total duration of a word or phrase, at the expense of syllable duration. This is exactly the opposite of what is predicted by a dichotomy of stress-timing versus syllable-timing.

The lack of compressibility of the syllable in some of the stress-timed languages is actually a consistent finding across studies. Crystal and House (1990) found that syllable duration has a quasilinear dependency on the number of phones in the syllable, and that the duration of stress groups has a quasilinear dependency on the number of syllables. Similar linearity patterns in English are found by van Santen & Shih (2000). For Swedish, Fant, Kruckenberg and Nord (1989) found the correlation between stress interval and number of phones to be as high as 0.94. Interestingly, in van Santen and Shih (2000), which examined segmental duration in both English and Mandarin, there is already some evidence that segments and syllables are more compressible in Mandarin than in English, which is consistent with the finding of Xu and Wang (2009). Given that temporal patterns indeed provide critical cues for the grouping of segments and syllables into larger units, it is actually a good question for future research why some of the languages labeled as stress-timed actually show greater linearity of duration as a function of segments and syllables than some other languages that are supposed to be syllable timed, such as Mandarin.

Another often-reported boundary marking cue is pitch reset (Ladd, 1988; Swert, 1997). Pitch reset is defined as the difference in pitch across a boundary. But it is usually measured in terms of the amount of drop, or the lack thereof, of F_0 peaks or top lines across the adjacent units (Ladd, 1988; Swerts, 1997). Swerts (1997) found that pitch reset measured this way was related to the perceived boundary strength in Dutch, but the correlation was only 0.35. A major factor of uncertainty in this regard is that pitch is already used extensively for lexical contrast, focus and modality, as discussed in previous sections. Thus there is a question of how much pitch is also used as a boundary cue. Wang, Xu and Ding (2018) examined the interaction between focus and boundary marking, and found that focus was signaled mainly by PFC, which can occur even across phrase breaks with silent pauses, whereas boundaries were mostly signaled by duration adjustments. The involvement of F_0 in boundary marking was only in terms of lowering of phrase-final F_0 minima and raising of phrase-

initial F_0 minima at relatively strong boundaries, i.e., those with silent pauses.

The interaction of focus and boundary marking brings us to the relationship between syntactic/semantic and prosodic structure (Ladd, 2008). Much work has attempted to account for the mismatches between prosodic structure and syntactic/semantic constituency (Kratzer & Selkirk, 2007; Selkirk, 1980; Steedman, 2000). From a functional perspective, these mismatches are not really an issue, as what is important is how utterances are divided into smaller chunks for ease of comprehension. The chunking can be marked by syntax, prosody, or both. Evidence for this can been seen in a recent study that examined how listeners detect boundaries (Buxó-Lugo et al., 2016). The findings are that a) syntactic position predicted whether listeners reported hearing boundaries, b) acoustic cues had a stronger effect for boundaries at syntactically unlicensed locations, and c) listeners report boundaries at licensed locations even when acoustic cues are weak. Thus both syntax and prosody provide useful cues for dividing utterances into smaller chunks, although there is also an incomplete division of labor. From this perspective, prosodic boundary marking is not for the sake of forming a prosodic structure for its own sake, but to either enhance a boundary already syntactically marked, or to provide boundary cues where syntax cues are ambiguous or absent (Allbritton, McKoon & Ratcliff, 1996; Lehiste, 1973; Price et al., 1991; Speer, Kjelgaard & Dobroth, 1996).

3.3. Paralinguistic prosody

Beyond post-lexical linguistic functions, there is also a rich set of paralinguistic meanings conveyed by prosody that are related to emotion, attitude, and vocal traits. Emotion- and attitude-related prosody is known as affective prosody, emotional prosody, or vocal expression of emotions and attitudes. The vocal-trait-related prosody does not even have a formal name, but cover things as variable as vocal attractiveness, charisma, dominance, sarcasm, irony, idiosyncratic prosody, etc. Though highly diverse, all these areas may be interconnected by a common mechanism first proposed for vocal calls in non-human animals by Morton (1977). This mechanism is extended to human speech through sound symbolism by Ohala (1983, 1984). Based on the analysis of calls by hundreds of bird and mammal species, Morton (1977) proposed that the calls are designed to mimic the acoustic properties of a large or small body size as a means to influence the behavior of the call receiver. When an animal is being aggressive, hostile, or is ready to attack, it would use calls with a low pitch and harsh quality to project a large body size in order to dominate the receiver. When being submissive, fearful or friendly, it would use calls that are high-pitched and tone-like to project a small body size in order to appease the receiver. In the latter case, the acoustic properties may also mimic infant vocalization so as to elicit a parental instinct. Ohala (1984) extended Morton's theory to human vocalization, and also added a third acoustic dimension: vocal tract length. He further proposed that the smile is actually an expression derived from a strategy to shorten the vocal tract during vocalization by retracting the corners of the lips. In addition, he proposed that body size projection is also what drove the sexual dimorphism, whereby men's larynges are lower and vocal folds longer than women's for the sake of projecting a larger body size in competition with other males for attracting female mates.

The Morton-Ohala hypothesis of body size projection has been most successfully applied in research on sexually related vocal attractiveness. It has been shown that female listeners favor men's voice with lower pitch and more condensed formants (hence longer vocal tract) (Collins, 2000; Feinberg et al., 2005; Xu et al., 2013b), while male listeners prefer female voice with higher pitch and more dispersed formants (Collins & Missing, 2003; Feinberg et al., 2008; Xu et al., 2013b). Xu et al. (2013b) further found, however, that in terms of voice quality, both male and female listeners prefer the same from the opposite sex: *breathy* rather than *pressed* voice. The breathiness nudges the voice toward a pure tone, a quality associated with friendliness and non-aggression (Morton, 1977).

Thus for an attractive male voice, intriguingly, the breathiness seems to soften the aggressiveness associated with low pitch and high formant density while preserving their association with masculinity (Xu et al., 2013b).

There is much less research applying the Morton-Ohala theory of body size projection to emotional prosody, however. A major difficulty seems to be that it is hard to find clear evidence from emotional speech through acoustic analysis (Scherer & Bängziger, 2004). This is presumably due to the heavy multidimensionality of emotional prosody in general, and the severe confound between pitch and loudness in particular, as will be discussed later. A method that is able to circumvent this difficulty is to use a perception paradigm in which the stimuli are generated either by direct speech synthesis or resynthesis of natural speech with manipulation of relevant acoustic dimensions, as suggested by Scherer and Bängziger (2004). When carried out by a number of studies, the paradigm was shown to be highly effective. Chuenwattanapranithi et al. (2008) found, with articulatorily synthesized (Birkholz, Jackèl & Kröger, 2006) isolated Thai vowels, anger and joy are associated with lowered or raised F_0 and lengthened or shortened vocal tract, respectively. Xu & Kelly (2010) replicated these findings with resynthesized British English numerals originally produced by a male speaker. The manipulation of formant dispersion (Fitch, 1994) was done by changing the spectral density with the PSOLA algorithm (Moulines & Charpentier, 1990). Similar findings were further made in Xu et al. (2013b), using a full sentence synthesized with an updated articulatory synthesizer capable of generating different voice qualities (Birkholz, Kröger & Neuschaefer-Rube, 2011).

With the three specific acoustic parameters, namely, pitch, formant dispersion and voice quality, body size projection is a kind of dimensional approach to emotional prosody. But it is very different from the much better known dimensional theories of emotions (Russell, 1979, 2003; Schlosberg, 1954). Those theories are based on a depiction of how we *subjectively feel* when being emotional: do we feel pleasant or unpleasant? How emotionally aroused are we? Or do we feel that we are in control or want to approach or avoid the hearer? These are known as the *valence*, *arousal/activation* and *control/approach-avoidance* dimensions, respectively (Carver & Harmon-Jones, 2009; Davidson et al., 1990). Of these dimensions, however, only arousal/activation has been shown to have relatively consistent acoustic correlates (Bachorowski, 1999; Scherer et al., 1991). No consistent acoustic correlates have been identified for the other dimensions, as reviewed by Mauss and Robinson (2009). The lack of acoustic correlates with the emotional dimensions except arousal/activation therefore contrasts with clear correlations between body-size-related acoustic parameters and happiness and anger (Chuenwattanapranithi et al., 2008; Xu et al., 2013a).

However, the body size projection dimensions may not be adequate to separate emotions other than anger and happiness, such as sadness, fear, disgust and surprise. Xu, Kelly and Smillie (2013a) proposed that a wider range of vocal emotions can be encoded by a set of *bio-informational dimensions* (BIDs) that consist of the body size projection dimension and three additional dimensions: *dynamicity*, *loudness* and *association* (Xu et al., 2013a). Like body size projection, the additional dimensions are also based on how they would influence the behavior of the receiver rather than how they are felt subjectively. *Dynamicity* is based on how vigorous the speaker wants to appear to the listener. *Audibility* controls the loudness of the vocalization, depending on whether it is beneficial for the vocalization to be heard only in close proximity. *Association* makes an associative use of sounds typically accompanying a non-emotional biological function in circumstances beyond the original ones, e.g., mirroring the sounds of vomiting to express disgust (Darwin, 1872). The BIDs would therefore allow an even richer multidimensional combination potentially used by many different emotions.

An initial test of BIDs was done by Xu et al. (2013a), using resynthesis of a full sentence in British

English to generate a multidimensional matrix of combinations of parameters that are associated with body size projection and dynamicity. The former is done by varying formant shift ratio and pitch median, and the latter by varying pitch range and speech rate. Among the main findings are the following.

1. In addition to a wide formant dispersion (hence short vocal tract length) as mentioned above, happiness also shows a large pitch range and fast speech rate, indicating high dynamicity.

2. Sadness show two highly distinct types: *depressed* and *grief-stricken*. While the depressed type shows similar characteristics as in most previous reports (Scherer, 2003; William & Stevens, 1972), with low values for virtually all parameters except for neutral formant dispersion, the grief-stricken type shows high pitch median, narrow pitch range, slow speech rate, and, surprisingly, low formant dispersion. The lengthened vocal tract as indicated by the low formant dispersion suggests that the crying vocalization, which is best seen in children, is for making a harsh demand rather than a gentle plea, as many parents would attest.

3. The most prototypical fear vocalization showed, also surprisingly, a narrow formant dispersion. This was interpreted as evidence that fear is not equal to submission, contrary to Morton's original hypothesis. But his principle still applies. That is, with a low formant dispersion, fear vocalization probably expresses a threat to fight on, especially when faced by a predator, as total submission would mean being eaten.

4. Finally, the best anger vocalization was heard from stimuli with 50 Hz median pitch, which sounded as if they were uttered with a vocal fry. This suggests that rough voice quality, which was only accidentally generated in the study, may be critical for expressing anger.

The contribution of voice quality was explored in a number of perception studies, using two methods. One is to coach a speaker to say a sentence with breathy, modal or tense voice. The other is to use VocalTractLab which then included a glottal modal for controlling voice quality (Birkholz, Kröger & Neuschaefer-Rube, 2011). The first method is found to be effective only in limited ways, as it is sometimes difficult to coach speakers to produce desired voice qualities consistently (Noble & Xu, 2011). The second method was better in terms of consistency, but its effectiveness varied with specific paralinguistic functions. Nevertheless, Xu et al. (2013b) was able to find, as mentioned earlier, that voice quality was a major determinant of vocal attractiveness, not only for female voice but also for male voice.

Thus the bio-informational dimensions, due to their ability to link specific acoustic parameters to theory-motivated hypotheses, enable the construction of testable predictions about various paralinguistic functions, including emotions, attitudes and vocal traits. They may also enable connection of previous findings that hitherto seem to be unrelated.

One major issue that is in need of investigation is about the audibility dimension proposed in Xu et al. (2013a). Production studies have mostly found high pitch to be characteristic of angry voice (Mauss & Robinson, 2009; Scherer, 2003; Williams & Stevens, 1972). Perceptual studies, however, have consistently found angry voice to be associated with a lower pitch than happiness (Chuenwattanapranithi et al., 2008; Xu & Kelly, 2010; Xu et al., 2013a, 2013b). Likewise, charismatic speech (Niebuhr, Voße & Brem, 2016; Rosenberg & Hirschberg, 2009) and speech that is heard to signal a high social rank (Ko, Sadler & Galinsky, 2014) are both found to have high F_0 even for male speakers, which again violates the prediction of body size projection. Here the potential confound is the natural correlation between F_0 and voice amplitude (Alku, Vintturi & Vilkman, 2002; Brumm & Zollinger, 2011): other things being equal, the higher the amplitude of the voice, the higher

the F_0. This is true of not only human voice, but also birdsongs, and is related to the well-known Lombard effect (Brumm & Zollinger, 2011; Nemeth et al., 2013). Thus it could be the case that the high F_0 of hot anger (Scherer, 2003) is closely related to a high loudness rather than for the sake of increasing pitch per se. The high loudness is already found in addition to high F_0 in the case of charismatic and high-rank speech (Ko et al, 2014; Niebuhr et al., 2016; Rosenberg & Hirschberg, 2009). But in general, the exact relationship between amplitude and F_0 in paralinguistic prosody is not yet clear, and so is in need of further research.

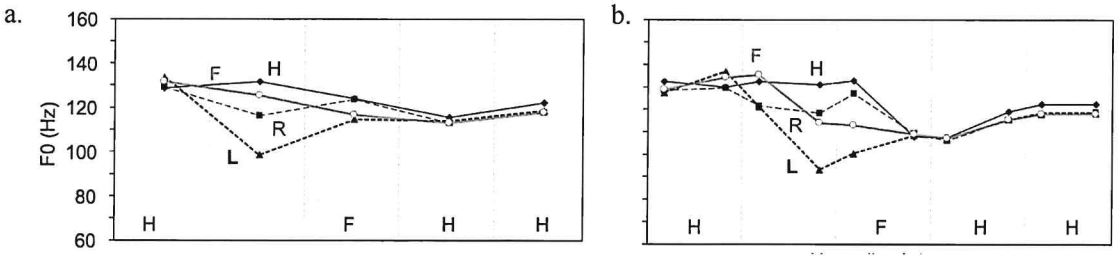
Finally, the paralinguistic functions are likely to frequently interact with the lexical and post-lexical linguistic functions, especially the latter. For example, the gradient pitch range variations related to the many modality types as shown in Figure 14 could well be partially due to the different paralinguistic connotations involved, e.g., uncertainty, surprise, incredulity and assertiveness. Also the prosodic difference sometimes found between contrastive and informational focus (Baumann et al., 2007; He et al., 2011; Sahkai, Kalvik & Mihkla, 2013) could also be related to their paralinguistic connotations. The study of detailed interactions between linguistic and paralinguistic functions may therefore lead to exciting new findings in further research.

4. A methodological note

The multiple degrees of separation, mentioned in 3.2, between prosodic functions and their underlying properties means that it is vital to apply methodological techniques that are effective in revealing potential causal relations. Some of the techniques have been illustrated throughout the chapter. This section is to highlight two such techniques with some elaborations.

1. Examining continuous prosody

While there has been an emphasis that the devil is in the fine phonetic details (Hawkins, 2003; Nolan, 1999), it is often not fully recognized that some of the most important details are in the continuous prosodic events such as F_0 contours. As shown in Figure 17a, if only a single F_0 measurement is taken from each syllable in an utterance (which is almost a common practice), the nature of the contextual tonal variation is not at all clear. Two to three measurements per syllable, as shown in Figure 17b and 17c, do improve the picture a bit further, but some important details are still missing. It is only when the contours appear fully continuous with 8 points per syllable, as shown in Figure 17d, does the evidence of asymptotic target approximation become obvious.



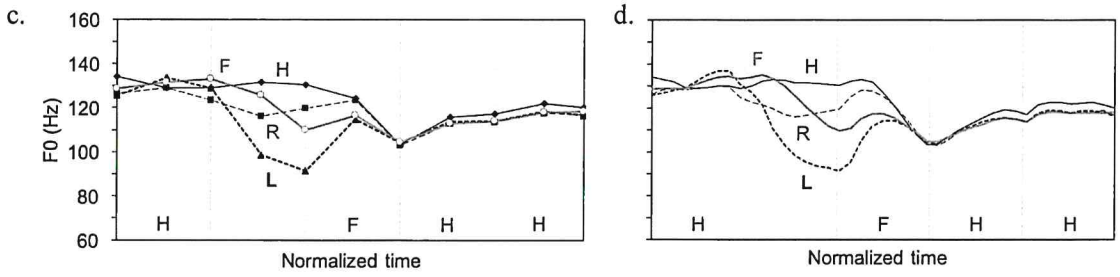
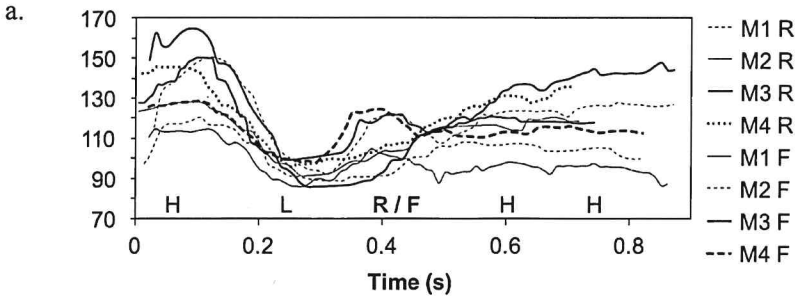


Figure 17. Plots of mean F_0 values taken from 5-syllable sentences in Mandarin. (a)-(c): 1–3 measurement point(s) per syllable, joined by straight lines. (d) 8 measurements per syllable. Data from Xu (1999).

2. Making minimal pair comparisons

The example in Figure 17 not only shows the importance of observing continuous prosodic events, but also illustrates the need for minimal-pair comparisons at the level of continuous trajectories. The concept of minimal pair comparison is familiar to anyone who does statistical comparison in their research. But minimal pair comparisons at the level of fully continuous trajectories are still relatively rare in the segmental (Gelfer, Bell-Berti & Harris 1989) as well as suprasegmental domains. A main reason is the difficulty in making the comparison. Figure 18a, for example, shows F_0 contours of two Mandarin tone sequences that differ only in the tone of the third syllable: Rising vs. Falling, produced by four male speakers. Because the utterances differ not only in F_0, but also in duration, it is hard to see the critical difference between the two sequences, although it is already a minimal pair design. In Figure 18b, the F_0 contours are time-normalized (i.e., with the same number of evenly spaced measurement points in each equivalent interval) with respect to the syllable and then averaged across 5 repetitions each by four speakers. The difference between the two sequences now becomes clear. In Figure 18c, the minimal-pair comparison is between neutral focus and medial focus. Again, the contrast is very clear thanks to the time normalization and averaging. It can also be seen clearly that due to the on-focus exaggeration of the Rising tone, the effect of the sharp rise does not fade away until half way into the following syllable. This graphical information tells us that, if we were to capture the PFC effect, the measurement points have to be taken after the strong carryover effect of the Rising tone is largely over.



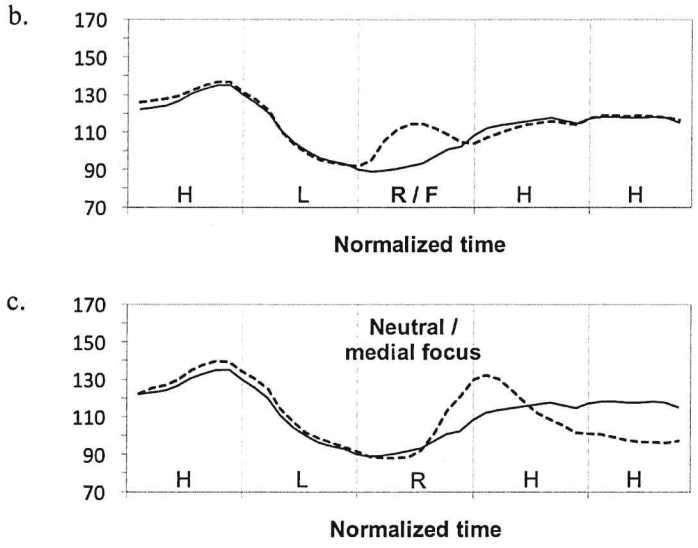


Figure 18. (a) F_0 contours of two Mandarin tone sequences differing in the tone of the third syllable, produced by four male speakers. (b) Mean F_0 contours time-normalized with respect to syllables, averaged across 5 repetitions each by four speakers. (c) Mean time-normalized F_0 contours averaged across 5 repetitions by four speakers each. Data from Xu (1999).

5. Concluding Remarks

This chapter has reviewed the literature on prosody, tone and intonation with the aim to address the question: how is it possible that a vastly rich set of communicative meanings can be all conveyed simultaneously through suprasegmental means? Past research has shown that articulatory mechanisms, lexical contrasts, post-lexical linguistic functions and paralinguistic prosody all have their own unique domain-specific mechanisms. Recognizing the specific mechanisms turns out to be key to understanding the highly diverse phenomena in all these areas. It is also shown that the individual processes can all be integrated into a framework that allows simultaneous transmission of multiple communicative meanings in parallel. Numerous questions, many of which were raised by the mechanistic-functional approach, nevertheless await being answered, as mentioned throughout the chapter. Answering those questions in future research would no doubt significantly advance our understanding of the suprasegmental aspect of speech.

References

Abe, I. (1998). Intonation in Japanese. In *Intonation Systems -- A survey of twenty languages*. D. Hirst and A. Di Cristo. Cambridge: Cambridge University Press. pp. 360-375.

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Alku, P., Vintturi, J. and Vilkman, E. (2002). Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. *Speech Communication* **38**: 321–334.

Allbritton, D. W., McKoon, G. and Ratcliff, R. (1996). Reliability of prosodic cues for resolving

syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **22**(3): 714-735.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics* **40**(3): 351-373.

Arvaniti, A. and Ladd, D. R. (2009). Greek wh-questions and the phonology of intonation. *Phonology* **26**(01): 43-74.

Atkinson, J. E. (1978). Correlation analysis of the physiological factors controlling fundamental voice freuquency. *Journal of the Acoustical Society of America* **63**: 211-222.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science* **8**(2): 53-57.

Baumann, S., Becker, J., Grice, M. and Mücke, D. (2007). Tonal and articulatory marking of focus in German. In *Proceedings of The 16th International Congress of Phonetic Sciences*, Saarbrucken: 1029-1032.

Beckman, M. E. (1996). The parsing of prosody. *Language and Cognitive Processes* **11**: 17-67.

Birkholz, P., Jackèl, D. and Kröger, B. J. (2006). Construction and control of a three-dimensional vocal tract model. In *Proceedings of The 31st International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France: I-873-876.

Birkholz, P., Kröger, B. J. and Neuschaefer-Rube, C. (2011). Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. In *Proceedings of Interspeech 2011*, Florence, Italy: 2681–2684.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* **5:9/10**: 341-345.

Bolinger, D. (1978). Intonation across languages. In *Universals of human language, Phonology V. 2*. J. H. Greenberg: Stanford University Press pp. 471-523.

Bomhard, A. R. (2008). *Reconstructing Proto-Nostratic: Comparative Phonology, Morphology, and Vocabulary*. Leiden: Brill.

Botinis, A., Ambrazaitis, G. and Frid, J. (2014). Syllable structure and tonal representation: revisiting focal Accent II in Swedish. *Proceedings from FONETIK 2014 Stockholm, June 9-11, 2014 PERILUS XXIV, June 2014*: 65.

Botinis, A., Fourakis, M. and Gawronska, B. (1999). Focus identification in English, Greek and Swedish. In *Proceedings of The 14th International Congress of Phonetic Sciences*, San Francisco: 1557-1560.

Browman, C. P. and Goldstein, L. (1992). Targetless schwa: an articulatory analysis. In *Papers in Laboratory Phonology II: Gesture, segment, prosody*. R. Ladd: Cambridge University Press pp. 26-36.

Bruce, G. (1982). Developing the Swedish intonation model. *Lund University, Dept. of Linguistics Working Papers* **22**: 51-116.

Brumm, H. and Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour* **148**(11-13): 1173-1198.

Buxó-Lugo, A. and Watson, D. G. (2016). Evidence for the influence of syntax on prosodic parsing. *Journal of Memory and Language* **90**: 1-13.

Carver, C. S. and Harmon-Jones, E. (2009). Anger is an approach-related affect: evidence and implications. *Psychological Bulletin* **135**(2): 183.

Chaffe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In *Subject and Topic*. C. Li. New York: Academic Press pp. 25-55.

Chahal, D. (2003). Phonetic Cues to Prominence in Lebanese Arabic. In *Proceedings of The 15th International Congress of Phonetic Sciences*, Barcelona: 2067-2070.

Chao, Y. R., (1968). *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

Chen, L., Li, X. and Yang, Y. (2012). Focus, newness and their combination: processing of information structure in discourse. *PloS one* **7**(8): e42533.

Chen, L., Wang, L. and Yang, Y. (2014). Distinguish between focus and newness: An ERP study. *Journal of Neurolinguistics* **31**: 28-41.

Chen, M. Y. (2000). *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge, UK: Cambridge University Press.

Chen, S.-w., Wang, B. and Xu, Y. (2009). Closely related languages, different ways of realizing focus. In *Proceedings of Interspeech 2009*, Brighton, UK: 1007-1010.

Chen, Ying (2015). Post-focus compression in English by Mandarin learners. In *Proceedings of The 18th International Congress of Phonetic Sciences*, Glasgow, UK

Chen, Ying, Xu, Y. and Guion-Anderson, S. (2014). Prosodic realization of focus in bilingual production of Southern Min and Mandarin. *Phonetica* **71**: 249-270.

Chen, Yiya (2008). Revisiting the Phonetics and Phonology of Shanghai Tone Sandhi. In *Proceedings of Speech Prosody 2008*, Campinas, Brazil: 253-256.

Chen, Yiya. and Xu, Y. (2006). Production of weak elements in speech -- Evidence from f0 patterns of neutral tone in standard Chinese. *Phonetica* **63**: 47-75.

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B. and Maneewongvatana, S. (2008). Encoding emotions in speech with the size code — A perceptual investigation. *Phonetica* **65**: 210-230.

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour* **60**: 773–780.

Collins, S. A. and Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour* **65**(5): 997-1004.

Connell, B. (2017). Tone and Intonation in Mambila. *Intonation in African Tone Languages* **24**: 131.

Connell, B. and Ladd, D. R. (1990). Aspects of pitch realization in Yoruba. *Phonology* **7**: 1-29.

Cooper, W. E., Eady, S. J. and Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America* **77**: 2142-2156.

Crystal, T. H. and House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America* **88**: 101-112.

Cutler, A., Dahan, D. and van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech* **40**: 141-201.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London, England: John Murray.

Davidson, R. J., Ekman, P., Saron, C., Senulis, J. and Friesen, W. (1990). Emotional expression and brain physiology I: Approach/withdrawal and cerebral asymmetry. *Journal of Personality and Social Psychology* **58**: 330-341.

de Jong, K. and Zawaydeh, B. (2002). Comparing stress, lexical focus, and segmental focus: patterns of variation in Arabic vowel duration. *Journal of Phonetics* **30**: 53-75.

De Pijper JR, Sanderman AA (1994) On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. J Acoust Soc Am 96:2037–2047.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for ΔC. In *Language and language processing: Proceedings of the 38th linguistic colloquium (pp.). Piliscsaba 2003*. P. Karnowski and I. Szigeti. Frankfurt: Peter Lang pp. 231-241.

Diamond, J. and Bellwood, P. (2003). Farmers and their languages: the first expansions. *Science* **300**(5619): 597-603.

Diamond, J. M. (1998). *Guns, germs and steel: a short history of everybody for the last 13,000 years*: Random House.

Dohen, M. and Lœvenbruck, H. (2004). Pre-focal rephrasing, focal enhancement and post-focal deaccentuation in French. In *Proceedings of The 8th International Conference on Spoken Language Processing*, Jeju, Korea: 1313-1316.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* **11**: 51-62.

Dilley, L. C. and Brown, M. (2007). Effects of pitch range variation on f0 extrema in an imitation task. *Journal of Phonetics* **35**(4): 523-551.

Duanmu, S. (1994). Against contour tone units. *Linguistic Inquiry* **25**: 555-608.

Eady, S. J. and Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America* **80**: 402-416.

Eady, S. J., Cooper, W. E., Klouda, G. V., Mueller, P. R. and Lotts, D. W. (1986). Acoustic characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and Speech* **29**: 233-251.

Erickson, D., Honda, K., Hirai, H. and Beckman, M. E. (1995). The production of low tones in English intonation. *Journal of Phonetics* **23**: 179-188.

Eriksson, A. (2012). Aural/acoustic vs. automatic methods in forensic phonetic case work. In *Forensic Speaker Recognition*: Springer pp. 41-69.

Face, T. L. (2005). F0 peak height and the perception of sentence type in Castilian Spanish. *Revista internacional de lingüística iberoamericana* **3**: 49-65.

Fant, G. and Kruckenberg, A. (1996). On the quantal nature of speech timing. In *Proceedings of Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on.* IEEE: 2044-2047.

Fant, G., Kruckenberg, A. and Nord, L. (1989). Stress patterns, pauses, and timing in prose reading. *STL-QPSR* **1**: 7-12.

Feinberg, D. R., DeBruine, L. M., Jones, B. C. and Perrett, D. I. (2008). The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. *Perception* **37**: 615-23.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M. and Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behavior* **69**: 561-568.

Fitch, W. T. (1994). *Vocal tract length perception and the evolution of language*. Ph. D. Dissertation, Brown University.

Féry, C. and Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics* **36**(4): 680-703.

Fok-Chan, Y. Y. (1974). *A perceptual study of tones in Cantonese*. Hong Kong: University of Hong Kong Press.

Franich, K. (2015). The effect of cognitive load on tonal coarticulation. In *Proceedings of The 18th International Congress of Phonetic Sciences*, Glasgow, UK

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech* **1**: 126-152.

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

Gandour, J., Potisuk, S. and Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics* **22**: 477-492.

Gay, T. J. (1968). Effect of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America* **44**: 1570-1573.

Gelfer, C. E., Bell-Berti, F. and Harris, K. S. (1989). Determining the extent of coarticulation: effects of experimental design. *Journal of the Acoustical Society of America* **86**(6): 2443-2445.

Giovanni, A., Ouaknine, M., Guelfucci, B., Yu, P., Zanaret, M. and Triglia, J.-M. (1999). Nonlinear behavior of vocal fold vibration: the role of coupling between the vocal folds. *Journal of Voice* **13**(4): 465-476.

Goldsmith, J. A., (1990). *Autosegmental and Metrical Phonology*. Oxford: Blackwell Publishers.

Grabe, E. and Low, E. L. (2002). Durational Variability in Speech and the Rhythm Class Hypothesis. In *Papers in Laboratory Phonology 7*. C. Gussenhoven and N. Warner. The Hague: Mouton de Gruyter pp. 515-546.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language* **2**: 73-113.

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*: Cambridge University Press.

Gussenhoven, C. (2007). Types of focus in English. In *Topic and Focus: Cross-linguistic Perspectives on Meaning and Intonation*. C. Lee, M. Gordon and D. Büring. New York: Springer pp. 83-100.

Halle, M. and Stevens, K. N. (1971). A note on laryngeal features. Quarterly Progress Report, M.I.T. Research Laboratory of Electronics. **101:** 198-213.

Halliday, M. A. K. (1967). Notes on Transitivity and Theme in English: Part 1. *Journal of Linguistics* **3**(1): 37-81.

Hanson, H. M. and Stevens, K. N. (2002). A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn. *Journal of the Acoustical Society of America* **112**: 1158-1182.

Hanssen, J., Peters, J. and Gussenhoven, C. (2008). Prosodic Effects of Focus in Dutch Declaratives. In *Proceedings of Speech Prosody 2008*, Campinas, Brazil: 609-612.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics* **31**: 373–405.

He, X., Hanssen, J., van Heuven, V. J. and Gussenhoven, C. (2011). Phonetic implementation must be learnt: Native versus Chinese realization of focus accent in Dutch. In *Proceedings of Proceedings of the XVIIth International Congress of Phonetic Sciences*: 843-846.

Hollien, H. (1960). Vocal pitch variation related to changes in vocal fold length. *Journal of Speech and Hearing Research* **3**: 150-156.

Hombert, J. M., Ohala, J. and Ewan, W. (1979). Phonetic explanation for the development of tones. *Language* **55**: 37-58.

Honda, K., Hirai, H., Masaki, S. and Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in F0 Control. *Language and Speech* **42**: 401-411.

Honorof, D. N. and Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *Journal of the Acoustical Society of America* **117**: 2193-2200.

Horne, M., Strangert, E. and Heldner, M. (1995). Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. In *Proceedings of Proceedings ICPhS*: 170-173.

Hwang, H. K. (2012). Asymmetries between production, perception and comprehension of focus types in Japanese. In *Proceedings of Speech Prosody 2012*, Shanghai: 326-329.

Hyman, L. M. (1993). Register tones and tonal geometry. In *The Phonology of Tone*. H. v. d. Hulst

and K. Snider. New York: Mouton de Gruyter pp. 75-108.

Hyman, L. M. (2007). Universals of tone rules: 30 years later. *Tones and tunes* **1**: 1-34.

Ipek, C. (2011). Phonetic realization of focus with no on-focus pitch range expansion in Turkish. In *Proceedings of The 17th International Congress of Phonetic Sciences*, Hong Kong: 140-143.

Ishihara, S. (2002). Syntax-Phonology Interface of Wh-Constructions in Japanese. In *Proceedings of Tokyo Conference on Psycholinguistics 2002 (TCP 2002)*, Tokyo: 165-189.

Ishihara, S. (2011). Japanese focus prosody revisited: Freeing focus from prosodic phrasing. *Lingua* **121**: 1870-1889.

Jeon, H.-S. and Nolan, F. (2013). The role of pitch and timing cues in the perception of phrasal grouping in Seoul Korean. *The Journal of the Acoustical Society of America* **133**(5): 3039-3049.

Jin, S. (1996). *An Acoustic Study of Sentence Stress in Mandarin Chinese.* Ph.D. dissertation. The Ohio State University.

Katz, J. and Selkirk, E. (2011). Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. *Language* **87**(4): 771-816.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics* **3**: 129-140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* **59**: 1208-1221.

Ko, S. J., Sadler, M. S. and Galinsky, A. D. (2014). The Sound of Power: Conveying and Detecting Hierarchical Rank Through Voice. *Psychological Science*.

Kratzer, A. and Selkirk, E. (2007). Phase theory and prosodic spellout: The case of verbs. *The Linguistic Review* **24**(2-3): 93-135.

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica* **55**(3): 243-276.

Kristensen, L. B., Wang, L., Petersson, K. M. and Hagoort, P. (2013). The Interface Between Language and Attention: Prosodic Focus Marking Recruits a General Attention Network in Spoken Language Comprehension. *Cerebral Cortex* **23**(8): 1836-1848.

Kubozono, H. (2007). Focus and intonation in Japanese: Does focus trigger pitch reset. In *Proceedings of Proceedings of the 2nd Workshop on Prosody, Syntax, and Information Structure (WPSI2)*: 1-27.

Kügler, F. and Genzel, S. (2014). On the elicitation of focus – prosodic differences as a function of sentence mode of the context? TAL 2014. Nijmegen: 71-74.

Kügler, F. and Skopeteas, S. (2007). On the universality of prosodic reflexes of contrast: The case of Yucatec Maya. In *Proceedings of The 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany

Ladd, D. R. (1988). Declination "reset" and the hierarchical organization of utterances. *Journal of the Acoustical Society of America* **84**: 530-544.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge: Cambridge University Press.

Laniran, Y. O. and Clements, G. N. (2003). Downstep and high raising: interacting factors in Yoruba tone production. *Journal of Phonetics* **31**: 203-250.

Lee, A. and Xu, Y. (2012). Revisiting focus prosody in Japanese. In *Proceedings of Speech Prosody 2012*, Shanghai: 274-277.

Lee, A., Prom-on, S. and Xu, Y. (2017). Pre-low raising in Japanese pitch accent. *Phonetica* **74**(4): 231-246.

Lee, Y.-c. and Xu, Y. (2010). Phonetic Realization of Contrastive Focus in Korean. In *Proceedings of Speech Prosody 2010*, Chicago: 100033:1-4.

Lee, Y.-c., Wang, B., Chen, S., Adda-Decker, M., Amelot, A., Nambu, S. and Liberman, M. (2015). A crosslinguistic study of prosodic focus. In *Proceedings of Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE: 4754-4758.

Lee, Y.-C., Wang, T. and Liberman, M. (2016). Production and Perception of Tone 3 Focus in Mandarin Chinese. *Frontiers in Psychology* **7**(1058).

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America* **51**: 2018-2024.

Lehiste, I. (1973). Phonetic disambigation of syntactic ambiguity. *Glossa* **7**: 107-122.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics* **5**: 253–263.

Lehiste, I.; Peterson, G. E., 1961. Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America* 33, 419-425.

Li, Y. (2015). Tone sandhi and tonal coarticulation in Fuzhou Min. In *Proceedings of The 18th International Congress of Phonetic Sciences*, Glasgow, UK

Liberman, M. and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In *Language Sound Structure*. M. Aronoff and R. Oehrle. Cambridge, Massachusetts: M.I.T. Press pp. 157-233.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry* **8**: 249-336.

Liberman, M., Schultz, J. M., Hong, S. and Okeke, V. (1993). The phonetic interpretation of tone in Igbo. *Phonetica* **50**: 147-160.

Lin, H. and Wang, Q. (2007). Mandarin rhythm: an acoustic study. *Journal of Chinese Language and Computing* **17**: 127-140.

Liu, F. and Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin

intonation. *Phonetica* **62**: 70-87.

Liu, F., Xu, Y., Prom-on, S. and Yu, A. C. L. (2013). Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modeling. *Journal of Speech Sciences* **3**(1): 85-140.

Ma, J. K., Ciocca, V. and Whitehill, T. L. (2006). Effect of intonation on Cantonese lexical tones. *The Journal of the Acoustical Society of America* **120**(6): 3978-3987.

Mauss, I. B. and Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion* **23**(2): 209-237.

Mei, T.-L. (1977). Tones and tone sandhi in 16th century Mandarin. *Journal of Chinese Linguistics* **5**: 237-260.

Mixdorff, H. (2004). Quantitative tone and intonation modeling across languages. In *Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing**:** 137-142.

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist* **111**: 855-869.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication* **9**(5-6): 453-467.

Nakatani, L. H., O'connor, K. D. and Aston, C. H. (1981). Prosodic aspects of American English speech rhythm. *Phonetica* **38**: 84-106.

Nemeth, E., Pieretti, N., Zollinger, S. A., Geberzahn, N., Partecke, J., Miranda, A. C. and Brumm, H. (2013). Bird song and anthropogenic noise: vocal constraints may explain why birds sing higher-frequency songs in cities. In *Proceedings of Proc. R. Soc. B*. The Royal Society**:** 20122798.

Niebuhr, O., Voße, J. and Brem, A. (2016). What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice. *Computers in Human Behavior* **64**: 366-382.

Noble, L. and Xu, Y. (2011). Friendly Speech and Happy Speech – Are they the same? In *Proceedings of The 17th International Congress of Phonetic Sciences*, Hong Kong**:** 1502-1505.

Nolan, F. (1999). The devil is in the detail. In *Proceedings of The 14th International Congress of Phonetic Sciences*, San Francisco**:** 1-8.

O'Connor, J. D. and Arnold, G. F. (1961). *Intonation of Colloquial English*. London: Longmans.

O'Malley, M. H., Kloker, D. R. and Dara-Abrams, B. (1973). Recovering Parentheses from Spoken Algebraic Expressions. *IEEE Transaction on Audio and Electroacoustics* **AU-21**: 217-220.

Odden, D. (1995). Tone: African languages. In *The Handbook of Phonological Theory*. J. A. Goldsmith. Cambridge, MA: Blackwell pp. 444-475.

Ohala, J. J. (1983). Cross-language uses of pitch: an ethological view. *Phonetica* **40**: 1-18.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice.

Phonetica **41**: 1-16.

Ortega-Llebaria, M. and Prieto, P. (2011). Acoustic Correlates of Stress in Central Catalan and Castilian Spanish. *Language and Speech* **54**(00238309): 73-97.

Palmer, H. E. (1922). *English Intonation, with Systematic Exercises*. Cambridge: Heffer.

Patil, U., Kentner, G., Gollrad, A., Kügler, F., Féry, C. and Vasishth, S. (2008). Focus, word order and intonation in Hindi. *Journal of South Asian Linguistics* **1**: 55-72.

Pedersen, H. (1931). *The Discovery of Language: Linguistic Science in the Nineteenth Century. English translation by John Webster Spargo*. Bloomington, IN: Indiana University Press.

Peng, S.-h. (2000). Lexical versus 'phonological' representations of Mandarin Sandhi tones. In *Papers in Laboratory Phonology V: Acquisition and the Lexicon.* M. B. Broe and J. B. Pierrehumbert. Cambridge: Cambridge University Press pp. 152-167.

Petrone, C., Truckenbrodt, H., Wellmann, C., Holzgrefe-Lang, J., Wartenburger, I. and Höhle, B. (2017). Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists. *Journal of Phonetics* **61**: 71-92.

Pham, A. H. (2003). The Key Phonetic Properties of Vietnamese Tone: A Reassessment. In *Proceedings of The 15th International Congress of Phonetic Sciences*, Barcelona**:** 1703-1706.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation.* Ph.D. dissertation, MIT, Cambridge, MA. [Published in 1987 by Indiana University Linguistics Club, Bloomington].

Pike, K. L. (1945). *The Intonation of American English.* Ann Arbor: University of Michigan Press.

Pike, K. L. (1948). *Tone Languages: A technique for determining the number and type of pitch contrasts in a language. With studies in tonemic substitution and fusion.* Ann Arbor: University of Michigan Press.

Port, R. F., Dalby, J. and O'Dell, M. (1987). Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America* **81**: 1574-1585.

Price, P. I., Ostendorf, M., Shattuck-Hufnagel, S. and Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America* **90**: 2956-2970.

Prom-on, S., Liu, F. and Xu, Y. (2012). Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling. *Journal of the Acoustical Society of America.* **132**: 421-432.

Prom-on, S., Xu, Y. and Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America* **125**(1): 405-424.

Ramus, F., Nesporb, M. and Mehlera, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition* **73**: 265-292.

Redi, L. and Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics* **29**: 407-429.

Remijsen, B. and Ayoker, O. G. (2014). Contrastive tonal alignment in falling contours in Shilluk. *Phonology* **31**(03): 435-462.

Rialland, A. (2009). African "lax" question prosody: its realisations and its geographical distribution. *Lingua* **119**: 928-949.

Rialland, A. and Robert, S. (2001). The intonational system of Wolof. *Linguistics* **39**: 893–939.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics* **1**(1): 75-116.

Rosenberg, A. and Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Communication* **51**(7): 640-655.

Rump, H. H. and Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech* **39**: 1-17.

Russell, J. A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology* **37**(3): 345-356.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review* **110**: 145-172.

Sahkai, H., Kalvik, M.-L. and Mihkla, M. (2013). Prosody of contrastive focus in Estonian. In *Proceedings of Interspeech 2013*, Lyon, France**:** 315-319.

Saindon, M. R., Trehub, S. E., Schellenberg, E. G. and van Lieshout, P. H. (2017). When is a Question a Question for Children and Adults? *Language Learning and Development*: 1-12.

Salffner, S. (2010). *Tone in the phonology, lexicon and grammar of Ikaan*, PhD dissertation, SOAS, University of London.

Savino, M. (2012). The intonation of polar questions in Italian: Where is the rise? *Journal of the International Phonetic Association* **42**: 23-48.

Schafer, A. J., Speer, S. R., Warren, P. and White, D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research* **29**: 169-182.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication* **40**: 227-256.

Scherer, K. R. and Bänziger, T. (2004). Emotional expression in prosody: a review and an agenda for future research. In *Proceedings of Speech Prosody 2004***:** 359-366.

Scherer, K. R., Banse, R., Wallbott, H. G. and Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion* **15**(2): 123-148.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological review* **61**(2): 81.

Schröder, M. (2001). Emotional speech synthesis: a review. In *Proceedings of INTERSPEECH*, Aalborg, Denmark**:** 561-564. [Widely cited (362), but most of the parameters are wrong and cluless]

Schuh, R. G. (1978). Tone Rules. In *Tone: A linguistic survey*. V. A. Fromkin. New York: Academic Press pp. 221-256.

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook* **3**: 371–405.

Selkirk, E. O. (1980). *On prosodic structure and its relation to syntactic structure*: Indiana University Linguistics Club.

Shattuck-Hufnagel, S. and Turk, A. E. (1996). A Prosody Tutorial for Investigators of Auditory Sentence Processing. *Journal of Psycholinguistic Research* **25**(2): 193-247.

Siemund, P. (2001). Interrogative constructions. In *Language typology and language universals*. M. Haspelmath, E. König, W. Oesterreicher and W. Raible. Berlin: Walter de Gruyter. **2:** pp. 1010-1028.

Sityaev, D. and House, J. (2003). Phonetic and phonological correlates of broad, narrow and contrastive focus in English. In *Proceedings of The 15th International Congress of Phonetic Sciences*, Barcelona**:** 1819-1822.

Speer, S. R., Kjelgaard, M. M. and Dobroth, K. M. (1996). The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of Psycholinguistic Research* **25**: 249-271.

Steedman, M. (2000). Information Structure and the Syntax-Phonology Interface. *Linguistic Inquiry* **31**: 649-689.

Steever, S. B. (1987). Tamil and the Dravidian languages. *The world's major languages*: 725-746.

Sugahara, M. (2005). Post-focus prosodic phrase boundaries in Tokyo Japanese: asymmetric behavior of an f0 cue and domain-final lengthening*. *Studia Linguistica* **59**(2-3): 144-173.

Svec, J. G., Schutte, H. K. and Miller, D. G. (1999). On pitch jumps between chest and falsetto registers in voice: Data from living and excised human larynges. *Journal of the Acoustical Society of America* **106**: 1523-1531.

Swerts, M. (1997). Prosodic features at discourse boundaries of different length. *Journal of the Acoustical Society of America* **101**: 514-521.

Swerts, M. and Zerbian, S. (2010). Prosodic transfer in Black South African English. In *Proceedings of Speech Prosody 2010*, Chicago

Taheri-Ardali, M. and Xu, Y. (2012). Phonetic Realization of Prosodic Focus in Persian. In *Proceedings of Speech Prosody 2012*, Shanghai**:** 326-329.

Terken, J. and Hirschberg, J. (1994). Deaccentuation of words representing 'given' information: Effects of persistence of grammatical function and surface position. *Language and Speech* **37**(2): 125-145.

Thorsen, N. G. (1980). A study of the perception of sentence intonation — Evidence from Danish. *Journal of the Acoustical Society of America* **67**: 1014-1030.

Tiffany, W. R. (1980). The effects of syllable structure on diadochokinetic and reading rates. *Journal*

of Speech and Hearing Research **23**: 894-908.

Titze, I. R., (1994). *Principles of Voice Production.* New Jersey: Prentice Hall.

Tsay, J. and Myers, J. (1996). Taiwanese tone sandhi as allomorph selection. In *Proceedings of Proceedings of 22nd Meeting of the Berkeley Linguistics Society*: 394-405.

Umeda, N. (1975). Vowel duration in american english. *The Journal of the Acoustical Society of America* **58**(2): 434-445.

Vainio, M., Mixdorff, H., Järvikivi, J. and Werner, S. (2003). The production and perception of focus in Finnish. In *Proceedings of Proceedings of ICPhS 2003*

van Heuven, V. J. and de Jonge, M. (2011). Spectral and Temporal Reduction as Stress Cues in Dutch. *Phonetica* **68**(3): 120-132.

van Heuven, V. J. and Haan, J. (2000). Phonetic correlates of statement versus question intonation in Dutch. In *Intonation*: Springer pp. 119-143.

van Santen, J. P. H. and Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *Journal of the Acoustical Society of America* **107**: 1012-1026.

van Santen, J., Kain, A., Klabbers, E. and Mishra, T. (2005). Synthesis of prosody using multi-level unit sequences. *Speech Communication* **46**: 365-375.

Wagner, M. (2005). *Prosody and Recursion.* Ph.D. Dissertation, Massachusetts Institute of Technology.

Wang, B., Wang, L. and Qadir, T. (2011). Prosodic encoding of focus in six languages in China. In *Proceedings of The 17th International Congress of Phonetic Sciences*, Hong Kong: 144-147.

Wang, B., Xu, Y. and Ding, Q. (2018). Interactive prosodic marking of focus, boundary and newness in Mandarin. *Phonetica* **75** (1): 24-56.

Wang, L., Wang, B. and Xu, Y. (2012). Prosodic encoding and perception of focus in Tibetan (Anduo Dialect). In *Proceedings of Speech Prosody 2012*, Shanghai: 286-289.

Wang, W. S.-Y. and Li, K.-P. (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research* **10**: 629-636.

Warner, N. and Arai, T. (2001). Japanese Mora-Timing: A Review. *Phonetica* **58**: 1-25.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* **91**(3): 1707-1717.

Williams, C. E. and Stevens, K. N. (1972). Emotion and speech: Some acoustical correlates. *Journal of the Acoustical Society of America* **52**: 1238-1250.

Wong, Y. W. (2006). Contextual Tonal Variations and Pitch Targets in Cantonese. In *Proceedings of Speech Prosody 2006*, Dresden, Germany: PS3-13-199.

Wong, Y. W. (2006). Realization of Cantonese Rising Tones under Different Speaking Rates. In *Proceedings of Speech Prosody 2006*, Dresden, Germany: PS3-14-198.

Wu, W. L. and Chung, L. (2011). Post-focus compression in English-Cantonese bilingual speakers. In *Proceedings of The 17th International Congress of Phonetic Sciences*, Hong Kong: 148-151.

Wu, W. L. and Xu, Y. (2010). Prosodic Focus in Hong Kong Cantonese without Post-focus Compression. In *Proceedings of Speech Prosody 2010*, Chicago

Xu, C. X. and Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association* **33**: 165-181.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics* **25**: 61-83.

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica* **55**: 179-203.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics* **27**: 55-105.

Xu, Y. (2001). Fundamental frequency peak delay in Mandarin. *Phonetica* **58**: 26-52.

Xu, Y. (2009). Timing and coordination in tone and intonation--An articulatory-functional perspective. *Lingua* **119**(6): 906-927.

Xu, Y. (2011). Post-focus compression: Cross-linguistic distribution and historical origin. In *Proceedings of The 17th International Congress of Phonetic Sciences*, Hong Kong: 152-155.

Xu, Y. and Kelly, A. (2010). Perception of anger and happiness from resynthesized speech with size-related manipulations. In *Proceedings of Speech Prosody 2010*, Chicago

Xu, Y. and Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication* **57**: 181-208.

Xu, Y. and Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America* **111**: 1399-1413.

Xu, Y. and Wallace, A. (2004). Multiple effects of consonant manner of articulation and intonation type on F0 in English. *Journal of the Acoustical Society of America* **115, Pt. 2**: 2397.

Xu, Y. and Wang, M. (2009). Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics* **37**: 502-520.

Xu, Y. and Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* **33**: 319-337.

Xu, Y. and Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics* **33**: 159-197.

Xu, Y., 2005. Speech Melody as Articulatorily Implemented Communicative Functions. *Speech Communication* 46: 220-251.

Xu, Y., Kelly, A. and Smillie, C. (2013a). Emotional expressions as communicative signals. In *Prosody and Iconicity*. S. Hancil and D. Hirst. Philadelphia: John Benjamins Publishing Co. pp. 33-60.

Xu, Y., Lee, A., Wu, W.-L., Liu, X. and Birkholz, P. (2013b). Human vocal attractiveness as signaled by body size projection. *PLoS ONE* **8**(4): e62397.

Yang, Y. and Wang, B. (2002). Acoustic correlates of hierarchical prosodic boundary in Mandarin. In *Proceedings of Speech Prosody 2002, International Conference*

Yip, M. (1989). Contour tones. *Phonology* **6**: 149-174.

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Yip, M. (2004). Phonological markedness and allomorph selection in Zahao. *Language and Linguistics* **5**: 969-1001.

Zemlin, W. R. (1998). *Speech and Hearing Science: Anatomy and Physiology (4th edition)*. Needham Heights, MA: Allyn & Bacon.

Zerbian, S. (2010). Developments in the study of intonational typology. *Language and Linguistics Compass* **4**(9): 874-889.

Zerbian, S., Genzel, S. and Kügler, F. (2010). Experimental work on prosodically-marked information structure in selected African languages (Afroasiatic and Niger-Congo). In *Proceedings of Speech Prosody 2010*, Chicago: 100976:1-4.

Zhang, J. (2014). Tones, tonal phonology, and tone sandhi. *The handbook of Chinese linguistics*: 443-464.

Zheng, X. (2006). Voice quality variation with tone and focus in Mandarin. In *Proceedings of The 2nd International Symposium on Tonal Aspects of Languages*, La Rochelle, France: 139-143.

Zsiga, E. and Nitisaroj, R. (2007). Tone Features, Tone Perception, and Peak Alignment in Thai. *Language and Speech* **50**(3): 343-383.