In Courses on Speech Prosody. A. R. Meireles: Cambridge Scholars Publishing pp. 146-177.

SPEECH PROSODY — THEORIES, MODELS AND ANALYSIS

Yi Xu, PhD[1]

Modern study of speech prosody started almost as early as modern study of segmental aspect of speech (Cruttendon, 1997). Over the decades, many theories and models are proposed. While the diversity of approaches is a sign of creativity of the field, the situation could be confusing for readers who are new to the area. Even for seasoned researchers, if they have not given much thought to methodological issues, the key differences between the many approaches may not be immediately clear. This chapter offers an overview of the state of the art in prosody research mainly from a methodological perspective. I will first try to highlight the critical differences between the theories and models of prosody by outlining a way of classifying them along a number of dividing lines. I will then discuss a number of key issues in prosody analysis, with focus also mainly on methodological differences.

1. Theories and models of prosody

1.1 Types of prosodic theories and models — A three-way division

One of the greatest difficulties in studying prosody is what can be referred to as the lack of reference problem (Xu, 2011). That is, due to the general absence of orthographic representations of prosody other than punctuations, there is little to fall back on when it comes to identifying the prosodic units, whether in terms of their temporal location, scope, phonetic property or communicative function. For example, for the F_0 plot shown in Figure 1, it is hard to determine what the relevant prosodic units are: F_0

[1] University College London.

peaks and valleys, turning points, size of the F_0 movements, temporal scope of a continuous movement, or all of them, or none of them. Because of the lack of orthographic representation of the prosodic units, it is difficult to decide whether any of them should or should not be considered as the relevant units. This difficulty lies in the heart of most of the theoretical disputes in speech prosody. In fact, depending on the degree of awareness of this difficulty and the amount of effort to overcome it, a three-way division can be made across the prosodic theories: *linear* vs. *superpositional*, *formal* vs. *functional*, and *acoustic* vs. *articulatory*. Although not all prosodic theories can be neatly fitted into this division scheme, the criteria defined here would help raise awareness of the issues that I believe are critical for theoretical development in prosody research. In the following, I will first describe the nature of each of the three divisions and discuss how existing major prosodic theories fit into this division scheme in one way or another.

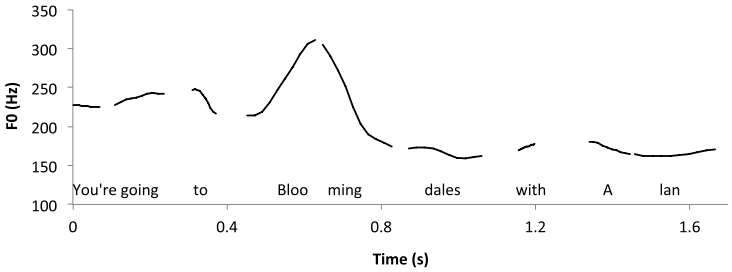


Figure 1: F0 track of "You're going to Bloomingdales with Alan" by a female American English speaker, with focus on "Bloomingdales". Data from Liu et al. (2013).

1.1.1 Linear versus Superpositional

Perhaps the most easily conceived prosodic theories are those that try to identify prosodic units directly from observed local acoustic patterns, such as F0 events. Because they conceptualize prosody as consisting of a string of discrete prosodic units, each occupying an exclusive temporal domain, these theories can be described as linear models. The exemplary ones include the British nuclear tone tradition (Crystal, 1969; O'Connor &

Arnold, 1961; Palmer, 1922; Wells, 2006), the AM (Autosegmental-Metrical) theory (Ladd, 2008), also known as the Pierrehumbert model (Pierrehumbert, 1980; Pierrehumbert and Beckman, 1988), and the IPO model (‘t Hart et al., 1990). Thus for the F0 contour shown in Figure 1, these models would have representations similar to those shown in Figure 2.

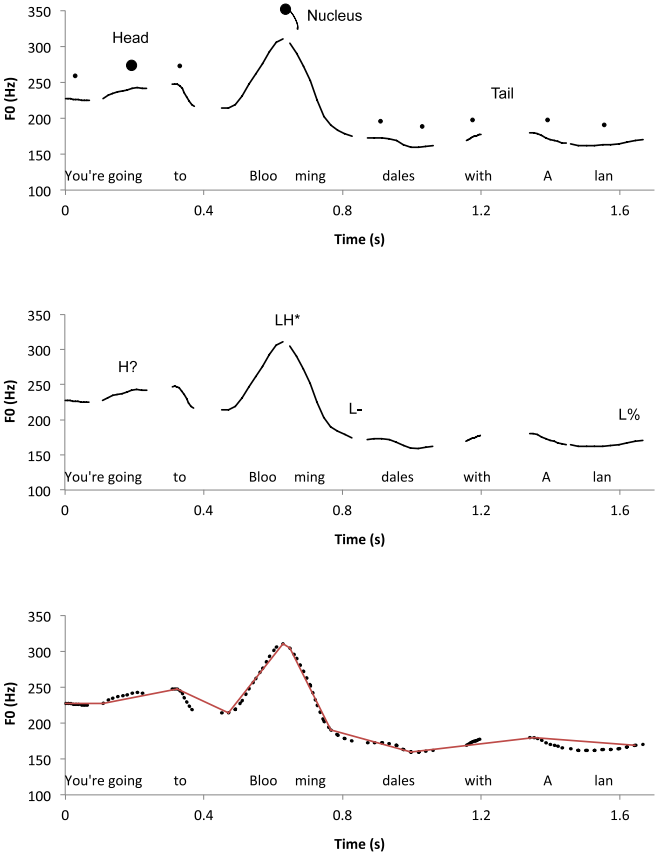


Figure 2: Schematic illustration of three linear models of prosody: Nuclear tone, AM and IPO, using the sentence in Figure 1 as an example.

All these approaches try to assign labels directly to local F_0 events, and allow only one prosodic unit for each temporal location. Sometimes, for example in the AM theory, some non-linear aspects of prosody are also recognized, such as overall pitch range, local pitch range (Ladd 2008; Pierrehumbert & Beckman, 1988), microprosody or segmental influences (Beckman 1995; Ladd 1995), and individual differences (Ladd 1995, 2008). However, it is insisted that only the linearly arranged events are contrastive, hence phonological (Ladd, 2008; Beckman, 1995), while the non-linear effects are gradient, and so are mostly phonetic or paralinguistic.

In contrast to the linear models are the superpositional models (Bailly & Holms, 2005; Fujisaki, 1983; Thorsen, 1980; van Santen et al., 2005). These models assume that surface F_0 contours are decomposable into *layers*, each consisting of a string of F_0 shapes, and the shapes of all the layers are added together to form surface F_0 contours. Figure 3 illustrates the conceptualization of the Fujisaki model (Fujisaki, 1983; Fujisaki et al., 2005). This model assumes two layers, corresponding to local shapes generated by accent commands, and global shapes generated by phrase commands. The two layers are added together on a logarithmic scale to form a smooth global surface F_0 contour.

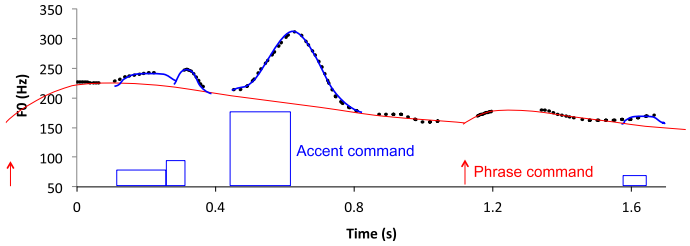


Figure 3: An illustration (i.e., not a real one) of how the F_0 contours generated by the Fujisaki model would fit the contour in Figure 1. The boxes represent the local accent command, while the arrows represent the global phrase commands.

1.1.2 Acoustic versus Articulatory

The second major division is based on whether a theory or model assumes that surface prosody is the result of direct acoustic manipulation or the product of articulatory control. For example, given the F_0 contour of an utterance, e.g., one of the contours in Figure 4a-d, one can easily find a close fit to it with either a high-order polynomial, or a low-order piecewise polynomial, such as a spline function, as is done in a number of studies (Andruski, 2004; Gandour, Tumtavitikul & Satthamnuwong, 1999; Grabe, Kochanski & Coleman, 2007; Hirst, 2005; Liu, Surendran & Xu, 2006). Some other models use non-polynomial curve fitting methods. Pierrehumbert (1981), for example, proposes a way to quantitatively implement the AM theory by fitting F_0 contours with linear and parabolic interpolations between adjacent F_0 turning points that are assumed to be associated with pitch accents. Taylor (2000) uses a tilt algorithm to characterize the shape of prominent F_0 peaks, and the shapes are then inter-linked by linear interpolation. The stylization approach in IPO is also largely a curve-fitting algorithm. In all these cases, reasonably good fit to the surface F_0 contours can be achieved for individual sentences, like those of Figure 4a-d. This is what may be called *ad hoc curve fitting*, in which the derived parametric representations are suitable for the specific utterances being fitted. However, different parameterizations may be needed for fitting different utterances. Thus for the curves in Figure 4a-d, at least four ad hoc curve fittings are needed, each for a specific tone sequence.

What is much more desirable is to achieve *predictive curve fitting*, i.e., to find invariant or quasi-invariant parametric representations of prosodic units to generate contours that can fit not only the original but also novel utterances. Predictive fitting is much harder than ad hoc fitting, of course, and the difficulty can be seen in Figure 4e, where the contours in Figures 4a-d are overlaid in a single graph. It can be seen that, as the tone of syllable 2 changes, the F_0 contours of the surrounding tones also vary extensively. Similar contextual variations can be also seen in Figure 4f, which differs from Figure 4e only in the tone of syllable 3. Predictive curve fitting would therefore require algorithms that can generate all the contextual tonal variants as those in syllable 3 in both Figure 4e and 4f.

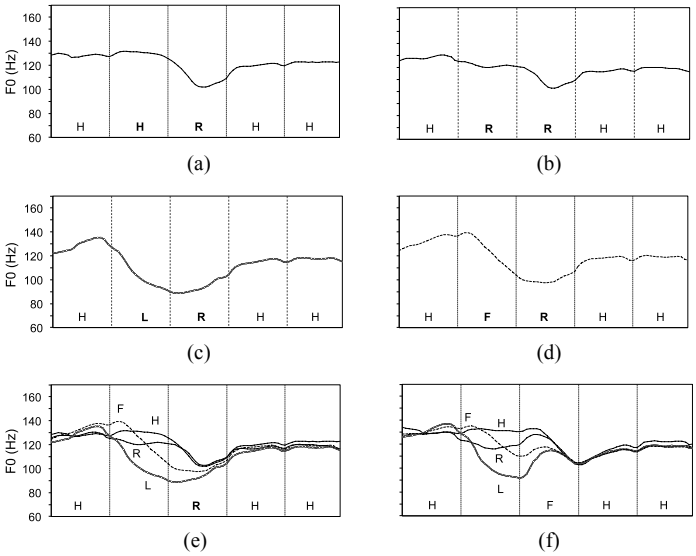


Figure 4: Mean time-normalized F_0 contours of 猫咪/迷/米/蜜 拿猫咪 [Kitty/Cat-fan/Cat rice/Cat honey picks up/sells kitty] produced by four males speakers of Beijing Mandarin (Xu, 1999). In all plots, vertical lines indicate syllable boundaries. Each contour is an average of 40 tokens said by four male speakers of Beijing Mandarin (5 repetitions by each). Adapted from Xu (1999).

Some acoustic-based models have been tested for predictive curve fitting. For example, the SFC model, which decomposes F_0 contours into superpositional Functional Contours without consideration of articulatory mechanisms, uses a neural-network controlled contour generator to choose from a large set of learned contours the ones that fit the functional profile (Bailly & Holms, 2005). Another approach, proposed by van Santen et al. (2005), models F_0 contours by decomposing pitch curves into re-combinable component curves, and combines them into units that fit into a

phonological hierarchy (segments, feet, phrases). Overall, however, most of the acoustic-oriented theories of prosody have not been systematically tested for predictive curve fitting, and so their ability in this respect is yet unknown.

In contrast to the acoustic-oriented models are articulatory-oriented models that attempt to take articulatory mechanisms of F_0 production into consideration in one way or another. These include the Fujisaki model (Fujisaki, 1983), the Stem-ML model (Kochanski & Shih, 2003) and the PENTA model (Prom-on et al., 2009; Xu, 2005; Xu & Prom-on, 2014). In the Fujisaki model, as illustrated in Figure 3, the shapes of local F_0 peaks and global F_0 trends are modeled as the on- and off-ramps of step and pulse responses of a second-order linear system. These responses are assumed to be associated with accent and phrase commands that are linguistically meaningful. Thus the commands, as the hypothetical underlying components of intonation, are different from the surface F_0 contours. And the latter are the product the underlying commands generated by the articulatory mechanism implemented in the model. The Stem-ML model simulates F_0 contours as deviations from underlying tonal templates under the influence of the surrounding tones (Kochanski & Shih, 2003). The surface F_0 contours are generated by a mechanism that compromises between maximum smoothness and full realization of the underlying tonal templates. The PENTA model, which will be detailed in the next section, simulates F_0 contours as syllable-synchronized laryngeal movements toward underlying pitch targets that are either static or dynamic (Xu, 2005). Thus all these models assume that surface F_0 contours result from certain articulatory mechanisms rather than from direct acoustic manipulations.

Just like in the case of acoustic-based models, for articulatory-based models there is also a question of whether and how well they can achieve predictive curve fitting. But only in a limited number of studies have attempts been made in this direction, e.g., Kochanski & Shih, (2003) for Stem-ML model, Prom-on et al. (2009) and Xu & Prom-on (2014) for PENTA model, and Raidt et al. (2004) for Fujisaki model.

1.1.3 Formal versus Functional

The final division between prosodic theories is about whether they assume *communicative functions* or *prosodic forms* as the defining properties of prosodic components. This division is not just about whether a theory recognizes the importance of communicative meanings in

prosody, as virtually all theories do. Rather, it is about whether a theory defines prosodic components primarily in terms of phonetic properties or communicative functions. The following quote from Pierrehumbert (1980:59) quite vividly illustrates the formal/functional divide:

> In the literature, one can distinguish two approaches towards the problem of establishing which intonation patterns are linguistically distinct and which count as variants of the same pattern. One approach attacks the problem by attempting to deduce a system of phonological representation for intonation from observed features of F0 contours. After constructing such a system, the next step is to compare the usage of F0 patterns which are phonologically distinct. The contrasting approach is to begin by identifying intonation patterns which seem to convey the same or different nuances. The second step is to construct a phonology which gives the same underlying representation to contours with the same meaning, and different representations to contours with different meanings.

Here the first approach defines components of prosody based on what appears obvious in the observed F_0 contours. In this way, the approach is largely *quasi-phonetic*, i.e., similar to the International Phonetic Alphabet. In fact, ToBI (Tones and Break Indices), which is partially derived from the AM theory, is known as an intonation transcription system (Silverman et al., 1992). Therefore, this approach can be also characterized as one of *analysis-by-transcription* (Xu, 2011). Theories that are the most typical of the formal approach include the British nuclear tone analysis, the AM theory (Gussenhoven, 2004; Ladd, 2008; Pierrehumbert, 1980), the IPO model (‘t Hart et al., 1990) and the tilt model (Taylor, 2000). The following quote from Pierrehumbert (1980:59) (which immediately follows the quote above), for example, states clearly that the prosodic categories of the AM theory are established based on a formal rather than functional principle:

> The work presented here takes the first approach, in fact, it stops at the first step in the first approach. While we hope that the system of phonological representation proposed here will be useful in investigating intonation meaning, we do not offer such a theory here. In some cases, rough descriptions of a meaning or usage of a particular contour are suggested. We include these only to help the reader picture what type of intonation is under discussion; there is no representation that they are a complete description of the meanings of the contour in question, nor that they are expressed in the correct terms of a theory which could provide such a complete description.

Thus it is clear that the proposal of the tonal categories in the AM theory is based on whether the *appearance* of the F_0 events, rather than the underlying *meanings*, are distinct from each other. In other words, although there is awareness of their associated meanings or functions, the

defining properties of the prosodic units are in terms their form rather than function. This is the fundamental difference between the formal and functional approaches to prosody. The principle is also behind later development of the Pierrehumbert model into the AM theory (Beckman & Pierrehumbert, 1986; Ladd, 1988; 2008), which tries to address those prosodic variations that exhibit superposition-like properties, such as pitch range variations. The proposed solution is to envisage a prosodic structure that organizes all the local components into a hierarchy (Beckman, 1995; Ladd, 2008). Such a hierarchy, again, is primarily defined in form rather than in function.

Similarly, some of the most classical theories of intonation, for example, Bolinger (1986, 1989), despite putting much emphasis on the pragmatic meanings of intonation, define intonational units like pitch accents based on directly observable intonational forms. This has led to the assertion that lexical stress in English does not have any clear prosodic correlates except when under focus (Bolinger, 1986, p. 14). In the case of Halliday (1967), the proposed meanings of theme and rheme are associated with prosodic units like accent, tonic nucleus, etc., which are established in the nuclear tone analysis tradition based on intonational form rather than function. In the AM theory, after the establishment of the form-based intonational phonology in Pierrehumbert (1980), an attempt is made to identify the meanings associated with phonological units such as H*, L*, etc. (Pierrehumbert & Hirschberg, 1990). But again, given that these units have already been established based on form, as mentioned above, prosodic meanings are treated only as secondary in defining the prosodic components.

In contrast to the formal theories and models, the functional approaches treat meaning and function as the defining properties of prosodic units. One clear example is the Superposition of Functional Contours (SFC) model (Bailly & Holm, 2005). SFC assumes that surface prosody results from superposition of multiple contours that each encodes a metalinguistic function. The metalinguistic functions considered in Bailly and Holm (2005) include segmentation, hierarchisation, emphasis and attitude that apply to units of variable sizes. In particular, they have considered functions like prosodic attitudes applied to sentences, dependency relations applied to syntactic constituents of read text or operands/operators of spoken math, cliticization typically applied to determiners and auxiliaries, narrow focus applied to words, and lexical tones applied to syllables in a tone language (Mandarin).

For both function- and form-oriented theories and models, however, there is an issue of how direct the link is between the prosodic events and

the meanings they carry. Some models assume that the link is very direct. SFC, for example, assumes that there is no intermediate representation of prosody (Bailly and Holm, 2005), and that prosodic events (in terms of overlapping multiparametric contours) directly encode deep phonological structures. Interestingly, a different kind of direct link is assumed by some form-oriented theories. According to Ladd (2008), it is widely held among linguists that intonational units directly carry morpheme-like meanings. One of the most elaborate cases is presented by Pierrehumbert and Hirschberg (1990), who propose that phonological units like pitch accents, phrase accents and boundary tones each directly carry a set of pragmatic meanings such as newness, salience, inter-phrasal relations, etc. But the assumption of such direct links has been questioned from within a similar framework of intonational phonology. Gussenhoven (2004:57) points out that in the English two-level calling intonation, which can be transcribed as H* H! in ToBI, neither H* nor H!, is meaningful by itself, and that the two-tone structure is like "one morpheme embodied in two phonological elements". As will be discussed in the next section, this kind of analogy to lexical morphemes can go even further, which may constitute a bridge that can eventually link the form-oriented and function-oriented theories of prosody.

1.2 PENTA — A quasi-superpositional, articulatory and functional model

In this section I present an outline of the parallel encoding and target approximation (PENTA) model (Xu, 2005) as an example of a quasi-superpositional, articulatorily based and function-oriented model. A schematic of PENTA is shown in Figure 5. The first block from the left represents communicative functions conveyed by speech. They are assumed to be parallel to each other, i.e., with no hierarchical relations, hence the key word 'parallel' in the name of the model. Such parallel encoding is similar in spirit to superposition, although it differs from it in terms of the assumed articulatory mechanisms. The second stack represents the 'encoding schemes' associated with the communicative functions. The schematization here makes it clear that the communicative functions *do not* directly control surface acoustics; rather, they are encoded by specific encoding schemes. It is assumed that the encoding schemes are either highly stylized and language specific or more gradient and universal, but their exact characteristics have to be empirically determined. The third block represents the articulatory parameters that are

controlled by the encoding schemes. These parameters in turn control the target approximation process represented by the fourth block. This is the mechanical process that generates surface F_0, which is done through the mechanism of *target approximation* (TA) as shown in the lower panel of the figure. In this mechanism, each syllable is assigned an underlying pitch target that is either dynamic (left dashed line in the lower panel) or static (right dashed line), and surface F_0 is the result of continuous articulatory approximations of successive pitch targets, and each target approximation movement is fully synchronized with the syllable associated with the target.

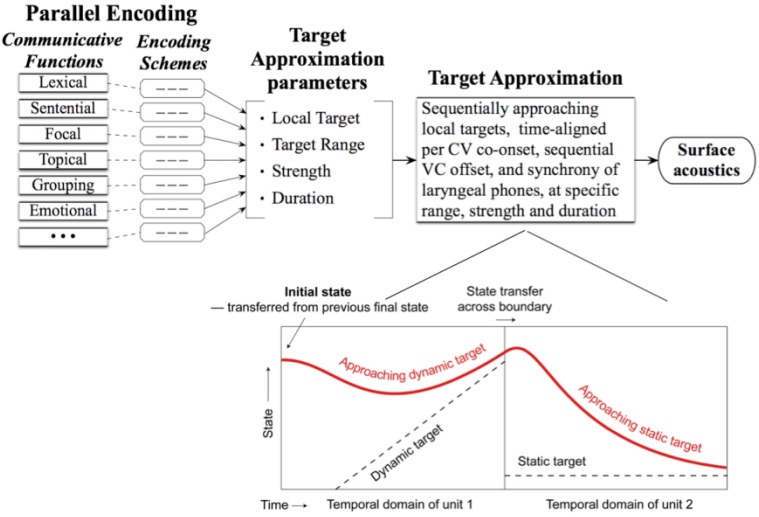


Figure 5: Upper panel: A schematic sketch of the PENTA model. Lower panel: The target approximation model of the articulation process. (Xu & Liu 2012; Xu & Wang 2001; Xu 2005)

Thus the core articulatory mechanism in PENTA is *syllable-synchronized target approximation*. It is based on the idea that human speech utterances are made up of linear sequences of syllables, and each syllable consists of articulatory movements toward a set of consonantal, vocalic and laryngeal targets (Xu & Liu, 2006). All the movements are synchronized with the syllable, except that syllable-initial consonants

complete their movements earlier than both the vocalic and laryngeal movements (Xu & Liu, 2006, 2012). With this mechanism, the only way to control the pitch of a syllable is to manipulate its underlying target, because speakers cannot realistically first articulate a string of pitchless syllables, and then impose an F_0 contour onto it afterwards, as is assumed in the genuine superposition models (Fujisaki et al., 2005; Bailly & Holm, 2005). The synchronization of target approximation movements with the syllable is based on the finding that the motor system is able to coordinate multiple movements only with full synchrony (Kelso, 1984; Kelso, Southard & Goodman, 1979; Mechsner et al., 2001). Syllable synchronization also means that when syllables become too short, both segmental and tonal targets may exhibit undershoot, sometimes very severely (Cheng & Xu, 2013; Xu & Wang, 2009). Also there are additional mechanisms that impose smaller yet noticeable influences on surface F_0 contours (Chen & Xu, 2006; Prom-on et al., 2012; Silverman, 1986; Whalen & Levitt, 1995).

According to the three-way division described above, therefore, PENTA can be described as *quasi-superpositional*, *articulatory* and *functional*. It is quasi-superpositional because it assumes that the observed prosody is the result of multiple communicative functions encoded in parallel rather than consisting of a single linear sequence of prosodic units. But it is "quasi-" rather than genuinely superpositional because it assumes that different communicative functions are encoded by modifying local pitch target parameters (height, slope and rate of approximation, cf. Xu, 2005 and Prom-on et al., 2009 for details), and so their effects on surface prosody is not linearly or logarithmically additive as assumed in pure superpositional models (Bailly & Holm, 2005; Fujisaki, 1983; van Santen et al., 2005). PENTA is an articulatory-based model because it assumes that surface F_0 contours result from syllable-synchronized target approximation, which is a hypothetical articulatory mechanism supported by increasing empirical evidence (Dilley, 2005; Gao & Xu, 2010; Niebuhr, 2007; Xu & Liu, 2006, 2012). Finally, PENTA is functional because it assumes that prosodic units are primarily defined by functions rather than by forms.

With regard to the link between form and meaning in prosody, PENTA, especially in its latest development (Liu et al., 2013; Xu et al., 2012), holds a position that differs from most other functional models. From its initial conception, PENTA assumes that meanings and prosodic forms are linked not directly, but through prosodic functions with specific encoding schemes that specify target parameters. More recently it is recognized that encoding schemes of prosodic functions can bear high

resemblance to lexical morphemes, in three critical ways. First, like lexical morphemes, each prosodic encoding scheme consists of multiple prosodic components, and these components are meaningless by themselves, but act jointly to mark both intra- and inter-functional contrasts. Second, similar to lexical morphemes, an encoding scheme of a prosodic function may have allomorph-like variants whose occurrences are conditioned by factors like location in sentence and interaction with other prosodic functions. Finally, similar to lexical morphemes, encoding schemes are language-specific and their patterns have likely historical sources. These encoding schemes differ from lexical morphemes in that they contrast prosodic functions that carry post-lexical meanings. It is therefore appropriate to give them a collective name: *prosopheme*.

One of the clearest examples of prosophemes is prosodic focus, whose function is to highlight one component against the rest of the sentence. Empirical studies have shown that focus is realized with not only specific pitch patterns, but also specific patterns of duration (Chen, 2007; Cooper, Eady & Mueller, 1985; Xu, 1999; Xu & Xu, 2005), intensity (Xu et al., 2012) and even voice quality (Sluijter & van Heuven, 1996). Also, focus is realized not only with prosodic patterns of the focused word itself, but also with *post-focus compression* (PFC) of pitch and intensity in many languages (see review in Xu et al., 2012). Furthermore, PFC has recently been found to be absent in many other languages (Xu et al., 2012). It is hypothesized that PFC as a special way of encoding focus is a feature inherited from a proto-language (Xu, 2011). Thus the encoding scheme of focus in languages like Mandarin and English are multi-componential, language specific, and with likely historical heritage, which is highly similar to lexical morphemes.

Another example is that, in American English, the underlying pitch target of a stressed syllable varies depending on whether the syllable is word final or non-final, whether the word is focused or not, and whether the sentence is a statement or yes-no question (Liu et al., 2013), as can be seen in Figure 7. Also can be seen in Figure 7 is that the F_0 of the post-focus syllables vary markedly depending on whether the sentence is a statement or question. In particular, post-focus F_0 in a question is raised well above the reference neutral-focus F_0. This pattern, however, is absent in Mandarin (Liu et al., 2013), as can be seen in Figure 6. Such a cross-linguistic typological difference is again in line with behavior of lexical morphemes, although more research is needed to further explore this phenomenon.

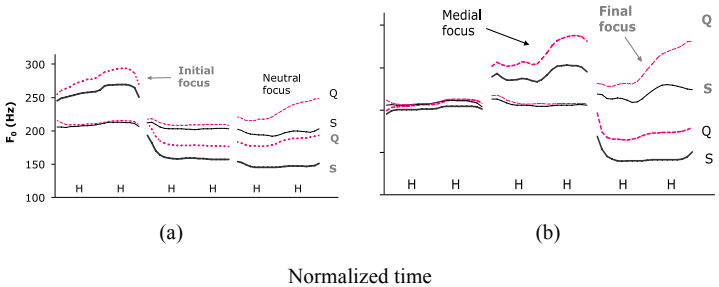


Figure 6. Mean F_0 contours of Mandarin sentence Zhāng Wēi dānxīn Xiāo Yīng kāichē fāyūn张威担心肖英开车发晕 [Zhang Wei is concerned that Xiao Ying may get dizzy when driving] spoken as either a statement or a question. On the left, either focus is on the sentence-initial word (thick lines), or there is no narrow focus (thin lines). On the right, focus is either sentence medial (thick lines) or sentence final (thin lines). The black solid lines represent statements, and the pink dashed lines represent questions. Data from Liu and Xu (2005).

In summary, the PENTA model shares the general spirits of many existing models of prosody, but differs from them in critical ways. With regard to surface prosody generation, it is in line with superpositional models in recognizing co-occurrence of multiple communicative functions in every local position, but it differs from many of them in insisting that the articulatory process of surface prosody generation has to be properly simulated. With regard to the link between meaning and form, PENTA is in line with the functional models in asserting that prosodic units should be primarily defined in function rather than form. But it also shares with the phonology oriented models in assuming that prosodic encoding of meanings is not direct in many cases, but through arbitrary constructs which can be viewed as prosodic morphemes, or prosophemes. Like lexical morphemes, these prosophemes typically consist of multiple phonetic properties, often follow language-specific arbitrary rules, and may have long historical etymologies.

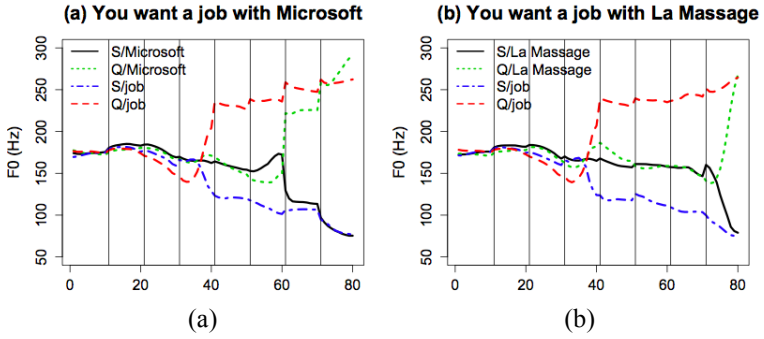


Figure 7. Mean F_0 contours of statements (S) and questions (Q) in American English. The word after "/" is focused. Data from Liu et al. (2013).

In summary, the PENTA model shares the general spirits of many existing models of prosody, but differs from them in critical ways. With regard to surface prosody generation, it is in line with superpositional models in recognizing co-occurrence of multiple communicative functions in every local position, but it differs from many of them in insisting that the articulatory process of surface prosody generation has to be properly simulated. With regard to the link between meaning and form, PENTA is in line with the functional models in asserting that prosodic units should be primarily defined in function rather than form. But it also shares with the phonology oriented models in assuming that prosodic encoding of meanings is not direct in many cases, but through arbitrary constructs which can be viewed as prosodic morphemes, or prosophemes. Like lexical morphemes, these prosophemes typically consist of multiple phonetic properties, often follow language-specific arbitrary rules, and may have long historical etymologies.

2 Analysis of prosody — A methodological perspective

Many of the issues concerned by the theories and models of prosody discussed so far are actually closely related to how the analysis of prosody is conducted, which will be the focus of this section. Since the early days of prosody research, many approaches to prosody analysis have been employed. They vary greatly in terms of the basic strategies and

underlying principles, and many of the disputes between theories and models of prosody, as reviewed in previous sections, are actually related to these methodological variations. Overall, there is a general trend toward improved methodological rigor and increased effectiveness in advancing knowledge of prosody. This is seen in the gradual progression through research strategies that may be characterized as *analysis by introspective transcription*, *analysis by acoustic transcription*, *analysis by hypothesis testing* and *analysis by modeling* (Xu, 2011). In the following sections, I will briefly describe each of these methodological approaches and offer some assessment of their effectiveness in advancing knowledge of speech prosody.

2.1 Analysis by introspective transcription

The early days of prosody research was dominated by descriptive methods that can be collectively characterized as analysis by introspective transcription. In this approach, symbolic representations of prosodic events are proposed based on the researcher's intuition and nonexperimental observation. One of the first such effort is Walker (1787), who proposed a tone marking system for English intonation that is not very different from the IPA annotations for lexical tones. The modern British intonation tradition is a continuation of this approach, with representative work by Palmer (1922), O'Connor and Arnold (1961), Halliday (1967), Crystal (1969), Cruttenden (1997) and Wells (2006). In this tradition, intonation is portrayed by a transcription system consisting of representations for prominences (usually by the size of successive dots corresponding to the stressed syllables) and contours (by curved lines, sometimes with arrow heads to indicate the direction of pitch movements), as illustrated in the top panel of Figure 2. Parallel to this tradition are transcription systems in America that emphasize tonal levels rather than tonal contours. This tradition can go back to Rush (1827), and the subsequent works include Pike (1945), Trager and Smith (1951) and Hockett (1958). A variant from this system is proposed by Bolinger (1986, 1989), who uses a transcription system that represents pitch contours by word-art like text arrangements. Just as influential as these transcription systems are works that adopt them as the basis for studying interfaces between prosody and syntax or pragmatics, for example, Chaffe (1974), Brazil et al. (1980) and Büring (2003). These works take the transcription-based prosody systems as if they were established facts.

An apparent issue with this general approach is that human introspection about prosody is not highly reliable. This is shown by the finding that human pitch awareness is not nearly as high as it is often assumed, especially when it comes to recognizing melodic events in prosody (Dankovicova et al., 2007). Thus both the establishment of the introspection-based prosodic categories and the proposed association of these categories to syntactic and pragmatic structure could have been affected by the imprecision of human introspection. Also, just as relevantly, a common practice among the works of this approach is that the evidence is typically presented in terms of examples, which, from an experimental point of view, are basically anecdotes without substantive verification or falsification (Popper, 1959).

2.2 Analysis by acoustic transcription

An apparent justification for analysis by introspective transcription in the early years is the lack of effective instrumentation for objectively observing prosodic events. This situation has been significantly improved over the years thanks to the availability of various hardware and software tools. This has led to the development of a new approach, namely, *analysis by acoustic transcription*. One of the best-known systems in this approach is Tones and Break Indices (ToBI) (Silverman et al., 1992). The system is developed based on the pitch accent representations proposed by Pierrehumbert (1980) and the boundary representations proposed by Price et al. (1991). A variant of ToBI has also been proposed by Breen et al. (2010). Compared to analysis by introspective transcription, the acoustic-based transcription approaches have the advantage of being based on something directly observable, not only to the transcribers, but also to other researchers. On the other hand, because the transcription system itself is derived from example-based work like Pierrehumbet (1980), Beckman and Pierrehumbert (1986) and Pierrehumbert and Hirschberg (1990), etc., the labeling exercise is seldom used as a way of questioning the original transcription system itself. Also just as importantly, as explained by Shattuck-Hufnagel & Turk (1996), the approach is not meant for developing a predictive system for prosody, but only as an initial step in that direction. In more recent research, analysis by acoustic transcription is incorporated into some empirical studies, in which the transcriptions are used as measurements and subjected to statistical analysis (e.g., Caspers, 2003; Grice et al., 2009; Mady & Kleber, 2010; Metusalem & Ito, 2008; Yoon, 2010). This is a welcoming step, but what is more important is to recognize the hypothetical nature of the categories in the transcription

systems themselves, which can be questioned if the experimental results fail to provide support. It is also important to compare these systems with other systems, especially those that do not follow the analysis by transcription approach.

2.3 Analysis by hypothesis testing

This is a very general approach that includes many different methods that can be described as experimental. What is common among these methods is the assumption that a theory is scientifically relevant only if it can produce hypotheses that can generate directly testable predictions (Popper, 1959). This is different from analysis by introspective transcription, for which hypotheses remain at the conceptual formation stage, and in many cases, further hypotheses are proposed on top of the initial ones which themselves have not yet been tested. It is also different from analysis by acoustic transcription, in which transcription systems that are hypothetical themselves are used as analysis tools, making it difficult to test the validity of the theory behind. In analysis by hypothesis testing, not only is there formation of general hypotheses, but also the hypotheses are made to derive predictions that can be empirically *falsified*. Neither the hypotheses nor the predictions are considered as corroborated until there is sufficient experimental support. And, even with strong corroborations, the prosodic categories themselves are not treated as part of the indispensible core of a theory, as in the case of the analysis by transcription approaches.
While these general principles are widely shared by researchers applying this approach, there are often disagreements on some specific, though critical issues. In the following, I will discuss, in particular, issues regarding *plausibility versus availability*, *existence versus usage*, *ecological validity versus predictability*, and *level of details*, with the goal to clarify some of the long-standing debates over methodology.

2.3.1 Plausibility versus availability

When it comes to determining the relation between form and meaning, one has to deal with the fact that there are many possible meanings that may be conceivably carried by prosody, and this is regardless of whether one assumes form or function as the defining properties of prosodic units. If every conceivable meaning had a corresponding prosodic

representation, there would simply be too many prosodic units. Thus it is unlikely that each and every conceivable meaning has a special prosodic representation. Hence, whether a meaning or function actually has a unique prosodic representation in a particular language can only be empirically established. A case in point is the widely assumed functional distinction between information (or presentational) focus and contrastive (or corrective) focus (Gussenhoven, 2007). Although the distinction is easily conceivable, the empirical findings so far have been mixed as to whether there are clear acoustic differences (Hanssen et al., 2008; He et al., 2011; Hwang, 2012; Sahkai et al., 2013; Sityaev & House, 2003). More importantly, there is as yet no evidence that listeners can identify them as two distinct functional types (Hwang, 2012). Also closely related to this case, Gussenhoven (2007) has proposed seven different types of focus in English, all of them based on conceptual plausibility. Again, empirical evidence is needed before we actually assume that there are indeed seven distinct types of prosodic focus in English or any other language.

In Xu (2011) an overview is provided of many plausible functions and their acoustic correlates reported by various studies. What is found in the survey is that the many plausible functions proposed over the years range widely in terms of whether distinct prosodic patterns have been found to be associated with them. The most distinctive ones include focus, modality and boundary marking. A number of other functions have been widespread in the literature but the empirical evidence has been mixed. One of the most extreme cases is contrastive topic. Its existence has been widely assumed in the syntax-pragmatics/prosody interface literature (Büring, 2003), yet no systematic experimental studies on its acoustic cues can be found in the literature (although see Ambrazaitis & Frid, 2012, which was published after that review).

One of the important reasons for the difficulty of finding clear acoustic correlates of some functions is the overlap in meaning between proposed functions. For example, prominence, newness and givenness all overlap with focus in one way or another. But there is some evidence that it is focus that has the most direct association with all of the acoustic cues previously proposed to be associated with the other functions (Liu et al., 2013; Wang & Xu, 2011; Xu, Xu & Sun, 2004). Also the meanings of topic, boundary marking and turn-taking seem to be at least partially overlapped. In each of these cases, as well as in other similar cases, there is a need for carefully designed studies to tease apart the overlapping hypothetical functions, and identify the ones that have the most consistent function-form associations.

2.3.2 Existence versus usage

Another issue that has been even less discussed explicitly is the difference between the *existence* of an encoding scheme (or prosopheme) in a language and the *circumstance of its usage*. Mixing the two is a likely source of confusion in prosody research. To understand this issue, it is helpful to draw from lexical morphology where this issue is easily understandable. There, it is obvious that the existence and exact form of a word or morpheme is a separate issue from the circumstances under which it is likely to be used. Usually the latter is much less certain than the former, because there are often alternative ways of conveying the same meaning. For prosody, a case in point is question intonation. In English and Mandarin, for example, experimental experience tells us that, whenever a naïve subject is presented with a sentence in text that ends with a question mark (whether or not the sentence has a polar question syntax), they would say the sentence with a rising intonation (together with its complex interaction with word stress and focus, cf. earlier discussion). This means that the rising intonation is part of their intonational repertoire and is the default pattern they would use when other factors are largely neutral. On the other hand, some studies have found that, in conversations, even syntactic polar questions are often said without rising intonation (Geluykens, 1988; Stivers, 2010). Thus some have argued that there is no one-to-one mapping between syntactic questions and rising intonation (Geluykens, 1988; Kohler, 2004). From a functional perspective, however, such a discrepancy only means that the syntactic structure is a sufficient, but not a necessary trigger of question intonation, because there are also other triggers, such as uncertainty, incredulity, information seeking, etc. It is therefore an empirical matter to identify all the conditions that trigger rising intonation.

Such triggering relation is probably the rule rather than the exception in prosody encoding. In other words, in order to study the encoding scheme of a function, it is sufficient to find a reliable trigger of the occurrence of the function, e.g., marking the end of a sentence with a question mark when making a laboratory recording. Once the encoding scheme is established, the knowledge can then be used to find out all the conditions that may trigger the occurrence of the function. In contrast, if our knowledge of an encoding scheme is still very vague, trying to identify its necessary trigger is likely to encounter many difficulties.

2.3.3 Ecological validity vs. predictive knowledge

Ecological validity refers to how closely the methods, materials and settings of an experimental psychology study approximate the real world (Brewer, 2000; Brunswik 1956). But the popular use of the term in speech science also often refers to how much of what is observed in a study is applicable to everyday speech. One extreme interpretation of the notion is that we should avoid studying speech, and particularly prosody, in the laboratory, and that preferences should be given to direct examination of spontaneous speech which is the closest to everyday speech (Hawkins, 2003; Local & Walker, 2005; Kohler, 2004). This kind of interpretation has often generated doubts among researchers about the value of their work using speech data collected in the laboratory. Curiously, however, the concern over ecological validity is usually directed against production studies only, while perception studies in the laboratory are much less criticized for lack of ecological validity. Like in the case of many other popular notions, there is a need to consider its logic thoroughly. It is true that speech, or at least some type of speech, produced in the laboratory can be somewhat "unnatural", especially when it is scripted. But the more crucial question is, how unnatural can the phonetics of such lab speech be? Are the subjects coached by the experimenters? Eve if they are, how much phonetic instruction can the experimenter really give the subjects? Put in this way, we can see that by and large subjects' own linguistic ability has to be responsible for the phonetic properties they produce in the laboratory. Even if in some cases when their speech can be described as unnatural, it is often because it lacks the rich variability that is observable in spontaneous conversations. But as any experimentalist would know, variability that is uncontrolled can only add noise to the data if its effect is random, or introduce confounds if the effect is biasing in one direction or another. And that is exactly the kind of difficulty one has to face in studying spontaneous speech, as is discussed in some detail by Beckman (1997) and Campbell (2004).

Turning now to perceptual studies, we can see that they can also be questioned about their ecological validity. That is, given that most perceptual experiments use speech stimuli that are digitally synthesized, how likely is it that the synthesized sounds actually occur in real life? This question can be put to even the most classical perceptual findings, such as those about categorical perception. In those experiments, typically one acoustic dimension is systematically varied while all the other dimensions are kept constant. But how likely are real speakers able to do the same

thing? If not, many of the stimuli heard by the listening subjects probably have never occurred in reality. Thus many of the perception findings, including even the classical ones, could be questioned if a stringent ecological validity standard is applied.

What is more important, however, is whether our research will lead to *predictive knowledge*, e.g., knowledge that is useful for the explanation, recognition and synthesis of prosody. Experimental investigations, by systematically controlling various factors, are designed to develop predictive knowledge, i.e., knowledge that is generalizable to other similar situations. However, also because of the need for systematic control, individual studies cannot examine all factors at once, and so each is necessarily limited in scope. For example, in a typical laboratory experiment, subjects are not asked to produce sentences with strong emotions, unless the study is about emotional expressions. Are those findings, then, still valid in cases where emotion is involved? This kind of issues should also be empirically resolved. The evidence so far is that functions like focus can be encoded in parallel with emotions (Xu et al., 2013). Bruce and Touati (1992) have demonstrated that prosodic patterns found in read speech in Swedish can also be found in spontaneous speech with rich emotions (political debate, radio listener call-in conversation, etc.). Of course, more such research is needed to further test the applicability of experimental findings to spontaneous speech. Likewise, studies that intend to directly embrace the richness of natural, spontaneous prosody should also go beyond only developing descriptive knowledge, and aim to establish predictive knowledge that can be applied in the explanation, recognition and synthesis of natural prosody.

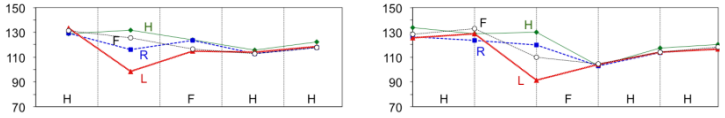
2.3.4 Level of details

Experimental control itself does not fully guarantee sure progress in our knowledge. Just as importantly, scientific understanding can be only as accurate as the level of detail we choose in our observations. In the case of prosody, if we choose to take measurements from only a limited number of points, such as at the F0 peak, valley, the center of a vowel, etc., although certain gross patterns can be observed, the causal relations among the contributing factors are likely to remain vague. In Figure 8a, for example, only one measurement is taken from the middle of each syllable. We can see that the largest differences occur in syllable 2, which carries four alternative tones. However, while statistics may show a significant difference across the four tones with this kind of measurement, many finer

differences are lost. In Figures 8b and 8c, as two or three measurements are taken from each syllable, more details start to emerge. But it is not until Figure 8d, where eight measurements are taken from each syllable, does the continuous dynamics of the F_0 contours become clearly visible.

So, as long as feasibility allows, observations and measurements should be as fine-grained as possible. Taking fine-grained measurements is time-consuming, of course, and an even greater problem is how to analyze them. One solution is to use time-normalization to allow averaging data points across repetitions by the same speaker or even across multiple speakers. The plots in Figure 8d, for example, are the averages of time-normalized F_0 contours over four male speakers, each producing five repetitions. Such time-normalized contours can now be easily obtained with ProsodyPro (Xu, 2013), a Praat script designed for large-scale systematic analysis of prosody. In addition to time-normalization, ProsodyPro also generates various measurements from the original, non-time-normalized contours, which can be used in statistical comparisons. Time-normalization and averaging also smooth out random variations unintended by the speaker, as well as individual differences, leaving only consistent variations due to tone and contextual tonal variations to be visible. From Figure 8 we can also see that time-normalization in Figure 8d is only a further extension of the coarser sampling shown in Figure 8a-8c, which are in fact also time-normalized. But the finer sampling allows us to see much more details, leaving little to guesswork. The detailed graphical comparisons enabled by time-normalization can also help us identify optimal measurements that best reflect the key differences between experimental conditions, and, just as importantly, avoid pitfalls.

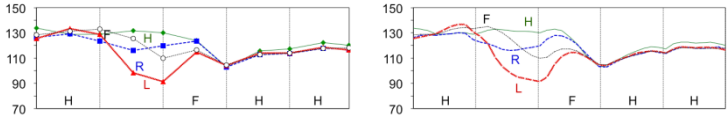
(to be continued)



(a) (b)

Normalized time

(continued)



(c) (d)

Normalized time

Figure 8: a.-c. Mean time-normalized F_0 of a five-syllable Mandarin sentence sampled at 1, 2, 3 and 8 samples per syllable.

2.4 Analysis by modeling

One of the ultimate goals of prosody research is to attain knowledge that is sophisticated enough to allow precise predictions of prosodic patterns. This in turn means that testing and improving the ability of theories to predict prosodic forms is an ultimate process of prosodic analysis. So, we are a full circle back to the starting point of this article, except that the focus now is on how to use computational modeling as a means of prosody analysis. The use of computational modeling as a way of expressing and testing scientific understanding has long been the common practice in fields like physics, chemistry, medicine and more recently biology, economy, etc. For speech science, computational models are well established in the case of acoustic theory of speech production (e.g., Fant, 1960; Stevens, 1998). However, for the dynamics of continuous speech, there has been only limited progress (e.g., Nam et al., 2012; Saltzman & Munhall, 1989). In this respect, prosody research is in fact somewhat more active than segmental research, probably because it is hard to study prosody without investigating events that change over time. However, as mentioned in 1.1.2, full predictive modeling is still rare even in prosody. More importantly, prosody modeling is often done for its own sake rather than as a tool of analyzing prosody, or for testing prosodic theories.

One of the first things that needs to be done in this approach is to convert conceptual theories into computational models that can generate continuous prosodic events whose details can be compared to those of real

speech. The comparisons can be in terms of both numerical evaluations such as root mean squared errors (RMSE) (Fujisaki et al., 2005; Kochanski et al., 2003; Prom-on et al., 2009; Raidt et al., 2004; Sun, 2002) and Pearson's correlation (Mixdorff & Jokisch 2001; Prom-on et al., 2009; Raidt et al., 2004; Sun, 2006), and perceptual evaluations of accuracy and naturalness (Sun, 2002, Prom-on et al., 2009, Ni et al., 2006). Those theories that are unquantifiable, therefore, cannot be tested this way, and thus cannot contribute much to the predictive knowledge of prosody.

The second thing is to make sure that models are tested for their ability to perform not only ad hoc curve fitting, but also predictive fitting, as discussed in 1.1.2. Predictive fitting, however, can be achieved at different levels. The highest level would be a system with human-like performance, i.e., starting from idea formation and finishing with production of fully natural and informative prosody. It will probably be a long time before anything close to that is achieved. The next best would be something akin to a concept-to-speech system (McKeown & Pan 2000; Young & Fallside, 1979), which, though also very tantalizing, seems to be also far from materialization anytime soon. The minimum level of predictive fitting would be that, given a set of utterances that are functionally marked, regardless of whether the category labels are derived from text or concepts, or determined by human labelers, can the modeling system generate prosodic forms that fit closely to those of the original? The modeling process can be also used as a means of hypothesis testing within a theory. For example, is it possible to answer specific questions like, is a particular communicative function prosodically encoded in a particular language? And if yes, how substantial is its contribution to surface prosody? What is the exact form of a particular prosodic encoding scheme?

3 Conclusion

The purpose of studying prosody is to gain a clear understanding of prosodic aspect of speech. How effectively this is done depends critically on the methodological approaches we adopt in our investigations. I have shown in this methodology-oriented overview of theories, models and analysis of prosody that there is a clear historical trend toward approaches that are hypothesis-driven, experimental-based, detail-sensitive, and modeling-oriented. Underlying this trend is drive to achieve predictive rather than just descriptive knowledge of prosody. With the wide availability of computer technology and software tools, achieving a solid

understanding of prosody is no longer just a dream, provided that we are ready to take bold and forward-looking steps and take full advantages of the technological advances.

References

Ambrazaitis, G. and Frid, J. (2012). The prosody of contrastive topics in Southern Swedish. In *Proceedings of FONETIK 2012*.

Andruski, J. and Costello, J. (2004). Using polynomial equations to model pitch contour shape in lexical tones: An example from Green Mong. *Journal of the International Phonetic Association* 34: 125-140.

Bailly, G. and Holm, B. (2005). SFC: a trainable prosodic model. *Speech Communication* 46: 348-364.

Beckman, M. E. (1995). Local shapes and global trends. In *Proceedings of The 13th International Congress of Phonetic Sciences*, Stockholm**:** 100-107.

Beckman, M. E. (1997). A typology of spontaneous speech. In *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Y. Sagisaka, N. Campbell and N. Higuchi. New York: Springer Verlag pp. 7–26.

Bolinger, D. (1986). *Intonation and its parts: melody in spoken English*. Palo Alto: Stanford University Press.

Bolinger, D. (1989). *Intonation and Its Uses -- Melody in Grammar and Discourse*. Stanford, California: Stanford University Press.

Brazil, D. M., Coulthard, M. and Johns, C. (1980). *Discourse Intonation and Language Teaching*. London: Longman.

Breen, M., Dilley, L. C., Kraemer, J. and Gibson, E. (2012). Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory* 8: 277-312.

Brewer, M. (2000). Research Design and Issues of Validity. In *Handbook of Research Methods in Social and Personality Psychology*. H. Reis and C. Judd. Cambridge: Cambridge University Press pp. 3-16.

Bruce, G. and Touati, P. (1992). On the analysis of prosody in spontaneous speech with exemplification from Swedish and French. *Speech Communication* 11: 453-458.

Brunswik, E. (1956). Perception and the representative design of psychological experiments: University of California Press.

Buitrago, N. (2013). *Types Of Focus In Spanish: Exploring The Connection Between Function And Realization*, PhD dissertation, Cornell University.

Büring, D. (2003). On D-Trees, Beans, and B-Accents. *Linguistics and Philosophy* 26(5): 511-545.

Campbell, N. (2004). Databases of expressive speech. *Journal of Chinese Language and Computing* 14(3-4): 295-304.

Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics* 31: 251–276.

Chaffe, W. (1974). Language and consciousness. *Language* 50: 111-133.

Chen, Y. (2006). Durational Adjustment under Contrastive Focus in Standard Chinese. *Journal of Phonetics* 34: 176-201.

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

Cheng, C. and Xu, Y. (2013). Articulatory limit and extreme segmental reduction in Taiwan Mandarin. *Journal of the Acoustical Society of America* 134: 4481-4495.

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.

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.

Cruttenden, A. (1997). *Intonation*. Cambridge: Cambridge University Press.

Crystal, D. (1969). *Prosodic Systems and Intonation in English*. London: Cambridge University Press.

Dankovicova, J., House, J., Crooks, A. and Jones, K. (2007). The Relationship between Musical Skills, Music Training, and Intonation Analysis Skills. *Language & Speech* 50: 177-225.

Dilley, L. C., Ladd, D. R. and Schepman, A. (2005). Alignment of L and H in bitonal pitch accents: testing two hypotheses. *Journal of Phonetics* 33: 115-119.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

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.

Fujisaki, H., Wang, C., Ohno, S. and Gu, W. (2005). Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command–response model. *Speech communication* 47: 59-70.

Gandour, J., Tumtavitikul, A. and Sattharnnuwong, N. (1999). Effects of speaking rate on Thai tones. *Phonetica* 56: 123-134.

Geluykens, R. (1988). On the myth of rising intonation in polar questions. *Journal of Pragmatics* 12(4): 467-485.

Grabe, E., Kochanski, G. and Coleman, J. (2007). Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech* 50: 281-310.

Grice, M., Baumann, S. and Jagdfeld, N. (2009). Tonal association and derived nuclear accents--The case of downstepping contours in German. *Lingua* 119(6): 881-905.

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.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English, Part II. *Journal of Linguistics* 3: 199-244.

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.

Hirst, D. J. (2005). Form and function in the representation of speech prosody. *Speech Communication* 46: 334-347.

Hockett, C. F. (1958). *A course in modern linguistics*. New York: MacMillan.

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

Kelso, J. A. S. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology: Regulatory, Integrative and Comparative* 246: R1000-R1004.

Kelso, J. A. S., Southard, D. L. and Goodman, D. (1979). On the nature of human interlimb coordination. *Science* 203: 1029-1031.

Kochanski, G. and Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication* 39: 311-352.

Kohler, K. J. (2004). Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions. In *From traditional phonology to modern speech processing*. G. Fant, H. Fujisaki, J. Cao and Y. Xu. Beijing: Foreign Language Teaching and Research Press pp. 205-214.

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

Ladd, D. R. and Terken, J. M. B. (1995). Modelling intra- and inter-speaker pitch range variation. In *Proceedings of The 13th International Phonetic Congress of Phonetics Sciences*, Stockholm**:** 386-389.

Liu, F. and Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62: 70-87.

Liu, F., Surendran, D. and Xu, Y. (2006). Classification of statement and question intonations in Mandarin. In *Proceedings of Speech Prosody 2006*, Dresden, Germany**:** PS5-25_0232.

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.

Local, J. and Walker, G. (2005). Methodological imperatives for investigating the phonetic organization and phonological structures of spontaneous speech. *Phonetica* 62: 120-130.

Mady, K. and Kleber, F. (2010). Variation of pitch accent patterns in Hungarian. In *Proceedings of Speech Prosody 2010*, Chicago

McKeown, K. R. and Pan, S. (2000). Prosody modelling in concept-to-speech generation: methodological issues. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 358(1769): 1419-1431.

Mechsner, F., Kerzel, D., Knoblich, G. and Prinz, W. (2001). Perceptual basis of bimanual coordination. *Nature* 414: 69-73.

Metusalem, R. and Ito, K. (2008). The role of L+H* pitch accent in discourse construction. In *Proceedings of Speech Prosody 2008*, Campinas, Brazil

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E. and Goldstein, L. (2012). A procedure for estimating gestural scores from speech acoustics. *The Journal of the Acoustical Society of America* 132(6): 3980-3989.

Ni, J., Kawai, H. and Hirose, K. (2006). Constrained tone transformation technique for separation and combination of Mandarin tone and intonation. *Journal of the Acoustical Society of America* 119: 1764-1782.

Niebuhr, O. (2007). The Signalling of German Rising-Falling Intonation Categories – The Interplay of Synchronization, Shape, and Height. *Phonetica* 64: 174-193.

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

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

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].

Pierrehumbert, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America* 70: 985-995.

Pierrehumbert, J. and Beckman, M. (1988). *Japanese Tone Structure*. Cambridge, MA: The MIT Press.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication*. P. R. Cohen, J. Morgan and M. E. Pollack. Cambridge, Massachusetts: MIT Press pp. 271-311.

Popper, K. (1959). *The Logic of Scientific Discovery (translation of Logik der Forschung)*. London: Hutchinson.

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: 405-424.

Raidt, S., Bailly, G., Holm, B. and Mixdorff, H. (2004). Automatic generation of prosody: Comparing two superpositional systems. In *Proceedings of Speech Prosody 2004*, Nara, Japan**:** 417-420.

Rush, J. (1827). *The philosophy of the human voice*. Philadelphia: J. B. Ltppincott & Co.

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

Saltzman, E. L. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology* **1**: 333-382.

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.

Shmueli, G. (2010). To explain or to predict? *Statistical Science* 25(3): 289-310.

Silverman, K. (1986). F0 segmental cues depend on intonation: The case of the rise after voiced stops. *Phonetica* 43: 76-91.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of The 1992 International Conference on Spoken Language Processing*, Banff**:** 867-870.

Stivers, T. (2010). An overview of the question–response system in American English conversation. *Journal of Pragmatics* 42(10): 2772-2781.

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.

Sluijter, A. M. C. and van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100: 2471-2485.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: The MIT Press.

Sun, X. (2002). *The determination, analysis, and synthesis of fundamental frequency*. Ph.D. dissertation, Northwestern University, 2002.

Surendran, D., Levow, G.-A. and Xu, Y. (2005). Tone Recognition in Mandarin using Focus. In *Proceedings of Interspeech 2005*, Lisbon, Portugal**:** 3301-3304.

't Hart, J., Collier, R. and Cohen, A. (1990). *A perceptual Study of Intonation — An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

Taylor, P. (2000). Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America* 107: 1697-1714.

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.

Trager, G. L. and Smith, H. L. (1951). *An outline of English structure*: Battenburg Press.

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

Walker, J. (1787). *The melody of speaking delineated*: (printed for the author, London 1787; reprinted by The Scolar Press, Menston. English Linguistics 1500–1800: No. 218, 1970).

Wang, B. and Xu, Y. (2011). Differential prosodic encoding of topic and focus at sentence initial position in Mandarin Chinese. *Journal of Phonetics* 39: 595-611.

Wells, J. C. (2006). *English intonation: an introduction*. Cambridge: Cambridge University Press.

Whalen, D. H. and Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics* 23: 349-366.

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

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

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46: 220-251.

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

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics* 38: 329-336.

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. (2011). Speech prosody: A methodological review. *Journal of Speech Sciences* 1: 85-115.

Xu, Y. (2013). ProsodyPro — A tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France**:** 7-10.

Xu, Y., Chen, S.-w. and Wang, B. (2012). Prosodic focus with and without post-focus compression (PFC): A typological divide within the same language family? *The Linguistic Review* 29: 131-147.

Xu, Y. and Liu, F. (2006). Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics* 18: 125-159.

Xu, Y. and Liu, F. (2012). Intrinsic coherence of prosodic and segmental aspects of speech. In *Understanding Prosody – The Role of Context, Function, and Communication*. O. Niebuhr. New York: Walter de Gruyter pp. 1-26.

Xu, Y. and Prom-on, S. (2014). Toward invariant functional representations of variable surface F0 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 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., Lee, A., Wu, W.-L., Liu, X. and Birkholz, P. (2013). Human vocal attractiveness as signaled by body size projection. *PLoS ONE* 8(4): e62397.

Yoon, T.-J. (2010). Speaker consistency in the realization of prosodic prominence in the Boston University Radio Speech Corpus. In *Proceedings of Speech Prosody 2010*, Chicago

Young, S. J. and Fallside, F. (1979). Speech synthesis from concept: A method for speech output from information systems. *Journal of the Acoustical Society of America* 66: 685-695.