

SPEECH PROSODY: A METHODOLOGICAL REVIEW

XU, Yi*

University College London

This critical review is mainly concerned with methodological issues in prosody research, with the aim to highlight progress toward developing predictive knowledge about prosody. The review shows that there has been a steady progression in terms of methodological rigor as the field goes through major methodological trends that can be described as analysis-by-introspection, analysis-by-transcription, analysis-by-hypothesis-testing and analysis-by-modeling. All the major methodologies currently still co-exist and each still has its own merit. But all of them are evaluated in terms of their effectiveness in establishing knowledge that is generalizable. Finally, an emphasis will be made on the need to have much more linking and integration between different subareas of prosody research.

Keywords: analysis-by-introspection; analysis-by-transcription; analysis-by-hypothesis-testing; analysis-by-modeling; predictive knowledge; degrees of freedom.

1 Introduction

Speech conveys information not only through segmental sounds like vowels and consonants, but also through prosody, i.e., variations in fundamental frequency (also referred to as F_0 and pitch), duration, intensity and voice quality. Unlike in the case of segments, one of the greatest difficulties about prosody is what I would call the *lack of reference problem*. By reference I mean a pivot that serves as both a starting point of inquest and a point that one can comfortably fall back on. In segmental research, for example, word identity serves as such a reference, because it is consciously accessible whether the language under study has a writing system or even whether the human informant is literate. Thus we can confidently investigate the phonetic properties that distinguish one word from another. But words also give us a false sense of certainty, because their ease of access may lead to the assumption that what underlies the lexical contrast, namely, the phonemes, are also easily accessible to individual speakers. This has been shown not to be the case by the discovery of phonological awareness (Mattingly, 1972; Liberman et al. 1990). That is, awareness of the segments needs to be either introduced or enhanced through literacy education (Bentin et al., 1992; Mann & Brady, 1988). So, the existence of orthographic representations of a functionality does not guarantee its conscious awareness and hence easy observation. In the case of prosody, very little of its functionality is orthographically represented, except for the punctuations whose meanings are at best ambiguous. Thus the starting point of inquest of prosody is inevitably vague and arbitrary, and it is just as difficult to know for certain what to check against after an observation is made. The difficulty is further compounded by the fact that our pitch awareness is not nearly as high as is often assumed, especially when it comes to melodic events in prosody (Dankovicova et al., 2007). As a result, the identification of the prosodic functionalities and their phonetic realizations often has to be attempted at the same time, with limited help from our built-in introspective ability. This is probably why approaches to prosody have been so disparate. There is therefore an urgent need to examine prosody research from the perspective of methodology. This paper offers a brief review of the strategies that have been attempted over the last century or so, with the hope that this may help future research to achieve greater efficiency in finding real solutions.

There have been a number of reviews of prosody research, the most recent only a year ago (Wagner & Watson, 2010). The emphases of those reviews are different from what is covered here. Shattuck-Hufnagel and Turk (1996) and Wagner and Watson (2010) were both concerned mainly with prosodic structures as reflected by prominence and boundary strength and their phonological representations. Cutler et al. (1997) focused on whether and how prosody may help the perceptual processing of words, syntactic structures and discourse structures. Botinis et al. (2001) was quite comprehensive and covered some of the issues discussed in the current review. But it was meant to be a tutorial and hence was in general fairly noncritical. Also much has happened in the past 10 years in terms of new empirical data and theoretical development. The current review will focus on the progress over the years toward solving the lack of reference problem by trying to build up predictive rather than descriptive knowledge about prosody.

2 Analysis by transcription

Systematic examination of prosody can go back as early as Walker (1787), who proposed a tone marking system for English intonation which 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), Kingdon (1958), 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). Parallel to this tradition, there have been transcription systems in America that put greater emphasis on tonal levels rather than tonal contours. This tradition can go back to Rush (1827), and the subsequent work includes Pike (1945), Trager and Smith (1951), Hockett (1958) and Pierrehumbert (1980). One exception from this tradition is Bolinger (1986, 1989), who used a transcription system that represents pitch contours by word-art like text arrangements. Despite the differences in the proclaimed general goals (one mainly for teaching and the other for analysis), the two traditions share one thing in common, namely, they both try to understand intonation through *analysis-by-transcription*. This is despite the improvement over the years on the tools of transcript from primarily auditory detection to instrumental observation. The latest major development of this approach, which is still widely practiced, is the transcription system known as ToBI for TOnes and Break Indices (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).

The methodological motivation behind the analysis-by-transcription approach is probably well characterized by Shattuck-Hufnagel & Turk (1996:193): "Because current theories do not predict the precise prosodic shape that a particular utterance will take, it is important to determine the prosodic choices that a speaker has made for utterances that are used in an auditory sentence processing study." In other words, given the fact that our knowledge about prosody has not yet developed to the level of being able to make precise predictions, there still needs to be means of referring to whatever prosodic forms we observe. It is interesting to note, however, the discrepancy between this characterization and the assumption widely held within this approach that the transcribed categories such as pitch accents, boundary tones and phrase tones are *phonological* rather than *phonetic*, i.e., the transcriptions represent phoneme-like rather than allophone-like units. As explained in Pierrehumbert (1980:59), the establishment of these categories was not based on whether or not they could clearly distinguish meanings, but rather was the outcome of "attempting to deduce a system of phonological representation for intonation from observed features of F_0 contours." In an effort to link the form-based intonation categories proposed in Pierrehumbert (1980), Pierrehumbert & Hirschberg (1990) proposed a compositional theory of tune interpretation, according to which the pitch accents H^* and L^* are meaningful themselves, i.e., directly representing meanings like newness, salience, mutual belief, incredulity, etc. Interestingly, this again differs from the classic assumption that segmental phonemes like consonants and vowels are not directly meaningful, but serve only as building blocks of morphemes that are the smallest units of meaning.

In general, therefore, as explained by Shattuck-Hufnagel & Turk (1996), analysis-by-transcription is not meant to develop predictive knowledge per se, but only as an initial step in that direction. But because the transcription systems themselves are not developed from empirical data, the lack of reference problem mentioned above is not fully addressed. In more recent research, analysis-by-transcription is often incorporated into 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 seems to be a welcome step, but it may also be beneficial to combine the transcription analysis with other measurements.

3 Analysis by introspection

Analysis-by-introspection refers to assignment of prosodic categories to text based mainly on intuition, without experimental investigation or transcription of recorded utterances. This kind of approach is adopted mainly in theoretical studies of syntax-prosody or pragmatics-prosody interfaces, for example, Chaffe (1974, 1976), Brazil et al. (1980) and Büring (2006, 2007). An apparent issue with this kind of approach is that, as mentioned in the introduction, human introspection is not known to be highly reliable. Thus the assigned prosodic categories could well have been affected by the imprecision of the introspection. Note that from the perspective of hypothesis testing to be discussed next, there is nothing wrong with using introspection as a means of hypothesis formation. What is crucial is not to end there, and treat the introspectively derived categories as established ones.

4 Analysis by hypothesis testing — Experimental approaches

By a broad definition, virtually any research effort contains some experimental elements. This is true of even analysis by introspection. What distinguishes different approaches is how many of the essential elements of a typical experimental paradigm are adopted. In analysis by introspection, the process remains at the stage of hypothesis formation, and in many cases, further hypotheses are proposed on top of the initial hypothesis which itself has not yet been empirically tested. A more thorough experimental approach would consist of not only the formation of a general hypothesis, but also predictions derived from the hypothesis that can be tested against empirical data, thus possibly falsifying the hypothesis. Falsifiability is the hallmark of genuine scientific inquiry according to the Popperian view of science (Popper, 1959). The process of falsification, though highly discipline-dependent, requires many essential components that are by now widely accepted, and have mostly been built into various statistical models. In general, it is essential to be able to independently control all the factors that might potentially confound the effect of the factors being tested. Such control is imperative in both data collection and data analysis. Note that, unlike what is often perceived, experimental control is not about dictating what the subjects should do. Rather, it is about how to guarantee systematic identification and separation of the factors that may significantly contribute to what is under observation. Such guarantee needs to be achieved by clear delimitation of independent, dependent, and to-be-controlled variables in the study design, preferably in a form that permits the application of statistical analysis.

4.1 Functionality versus encoding

For research on prosody, an additional issue, as mentioned in the introduction, is how well a study manages to separate functionality from encoding. This issue is methodologically essential because it is directly relevant for the question of control. Ideally, a prosodic function should be defined in terms of both the kind of information it conveys and its prosodic forms of encoding, but it is vitally important that the two parts of the definition be clearly separated. On the one hand, one would look for prosodic functions that make contrasts that are defined in meaning rather than in form. On other hand, each operational function should have, by definition, forms that directly encode the contrast. So, it is an empirical question whether each hypothesized

function actually has a unique encoding scheme in a particular language. [Table 1](#) (p. 110-111) shows a list of plausible functions and their acoustic correlates as reported by various studies.

As can be seen in [Table 1](#) (p. 110-111), focus and boundary marking have both attracted a large amount of research effort, and clear acoustic cues have been identified. For focus, increasingly clear evidence shows that its cue is not only in the focused items themselves, but also in the post-focal components, which typically show reduced pitch range and intensity. But such *post-focus compression* (PFC) is found only in some of the languages, while many other languages show no evidence of PFC. The cause of such uneven distribution is currently under active investigation (cf. Xu, 2011 for a brief overview). Also related to focus are *prominence* and *newness*, two potential functions that have been widely recognized. Prominence has sometimes been viewed as synonymous to focus, but often it is assumed to be free of any specific linguistic functions. Such freedom, however, leaves room for confounding between better defined functions such as lexical stress and focus. The problem resulting from such confounding is highlighted by the recent controversial finding by Kochanski et al. (2005) that F_0 contributes little to prominence. In the case of newness/givenness, conceptually the contrast can be easily defined (Bock & Mazzella, 1983). However, when examined with other functions systematically controlled, it does not seem to show distinct prosodic correlates independent of focus and topic, other than a small durational effect, at least in Mandarin (Wang & Xu, in press). In the studies on other languages, newness is not genuinely separated from focus and therefore it is yet unknown whether it has unique cues of its own. Another focus-related issue is the different types of focus that have been widely assumed (Gussenhoven, 2007). Of the four studies shown in [Table 1](#) (p. 110-111), that have looked into different types of focus, only one has reported difference between narrow and contrastive focus (Avesani & Vayra, 2003, for Florence Italian). More research is therefore needed to further investigate whether there is distinct prosodic coding for different types of focus.

There have been remarkably many studies looking into issues related to boundary marking, and the list shown in [Table 1](#) (p. 110-111) is far from complete. Overall, the most consistent cue is in terms of duration, whether it is that of domain-final vowel, syllable or word, pause or domain-initial consonant. The emerging picture is that such duration-based boundary marking is applied across the board, from within a word to within a phrase to between phrases to between sentences, as can be seen in Table 2. Also shown in Table 2 is an increase in the number of cues as the units separated by the boundary become larger. Interestingly, as can be seen in the bottom two rows, when the unit is the sentence or larger, boundary cues seem to be joined by topic cues. This may suggest that the topic function could be a special case of the boundary function. This possibility can be explored in future research.

Table 2: Prosodic boundary cues reported in a number of studies. An increase in the number of cues can be seen as the boundary level becomes higher.

Boundary type	Prosodic cues
Within phrase	Duration (Gussenhoven & Rietveld 1992, Nakatani et al. 1981, White & Turk 2010, Xu & Wang 2009)
Between phrases	Duration, F_0 (Edwards et al. 1991, Gussenhoven & Rietveld 1992, Klatt 1975, Nakatani et al. 1981, Wagner 2005)
Between sentences	Duration, pause, F_0 (Berkovits 1993, 1994, Ladd 1988, Swerts 1997, Wang & Xu in press)
Between paragraphs (topic?)	Duration, pause, F_0 (Nakajima & Allen 1993, Lehiste 1975, Swerts 1997)

Within the topic function, there is a subtype known as contrastive topic, which is widely recognized among the syntacticians (Büring, 2003; Kiss, 2008; Krifka, 2008). It is assumed that contrastive topic is associated with a rising pitch pattern known as B-accent (Bolinger, 1965). However, judging from the pitch track provided in Büring (2003) which is cited from Jackendoff (1972), there is no clear evidence of rising F_0 . Also in the course of conducting this review, no systematic empirical data were found on contrastive topic. According to Büring (2007), the contrastive component of contrastive topic is similar to focus. At least for Mandarin, Wang & Xu (in press) found that contrastiveness in topic has no prosodic correlates independent of focus when both topic and focus are controlled. This issue nevertheless needs more investigation for other languages.

The prosody-syntax row in [Table 1](#) (p. 110-111) shows studies on how prosody and syntax influence each other and those on how prosody may aid comprehension in speech perception. The prosodic cues are rather diverse and often not clearly stated. The relation between prosody and syntax is nevertheless a very important issue, and there is a question that has rarely been asked before: Why should the two be matched or linked in the first place? That is, assuming that speech is about exchanging information, if a meaning is already syntactically or prosodically conveyed, why should it be also encoded in the other domain? Unless, of course, *redundancy of coding* is favored (Assmann & Summerfield, 2004). This possibility suggests that exploring cross-domain relation in terms of redundant coding is a direction worth pursuing in future research.

The last two rows of [Table 1](#) (p. 110-111) show that there have been many studies looking at emotions and attitudes. A multitude of cues have been reported for various emotions and attitudes and they are too diverse to be discussed in detail in this review. Interested readers may find a number of detailed literature reviews on vocal expression of emotions and attitudes, including Mauss and Robinson (2009), Murray and Arnott (1993), Scherer (2003) and Scherer and Bänziger (2004). In addition, a new line of research based on evolutionary mechanisms and cross-species comparisons has shown some interesting results (Chuenwattanapranithi et al. 1996; Morton, 1977; Ohala 1984, 1996; Xu & Kelly, 2010; Xu et al., forthcoming).

4.2 Production vs. perception

Among the experimental approaches, there is often a preference for either a production- or perception-oriented strategy. From a functional perspective, an operable function by definition must be contrastively encoded through production, and reliably decoded through perception. The production and perception of a function therefore must have evolved in tandem: only patterns that are perceptually distinct and articulatorily possible could have emerged and been maintained. It is therefore beneficial for both perception- and production-oriented studies to take a comprehensive view in their interpretation of the empirical data. In the following, I will highlight a few issues as an illustration of the importance of a comprehensive view.

Pitch contour stylization is a strategy to find a piecewise linear approximation of the F_0 curve that is perceptually indistinct from the original contours. It was developed in the 60-80s with the assumption that pitch movements that are interpreted as relevant by listeners in non-linguistic perceptual judgment tasks directly reflect the intention of the speaker ('t Hart et al. 1990). An important issue about this and other stylization approaches is generalizability. While piecewise linear approximations may be good enough for specific utterances, it may not be applicable to other utterances with different segmental and lexical compositions. This is because linear approximations have no built-in mechanisms to handle contextual variations, which alter F_0 height, movement direction and alignment of turning points (Gandour et al., 1994; Potisuk et

al., 1996; Xu 1999). These contextual variations have many different sources, as summarized in Xu (2006), and each need to be handled in terms of its own specific underlying mechanism.

Likewise, for any pattern found in production, there is also the question of perceptual relevance and sensitivity. A case in point is the many acoustic patterns found to be associated with focus in production studies. There has been some research on the perceptual relevance of focus-related pitch patterns (Botinis, 2000; Prom-on et al. 2009; Rump & Collier, 1996; Xu et al., 2004). Prom-on et al. (2009) also found that incorporating focus-appropriate duration changes further improves focus identification rate based on F_0 only changes. However, there has been virtually no research examining the perceptual relevance of intensity (Kochanski et al. 2005) and spectral balance (Sluijter & van Heuven, 1996).

Another pattern found in recent research is a rather consistent temporal alignment of F_0 turning points relative to segmental events, especially to syllable edges (Arvaniti & Garding, 2007; Atterer & Ladd, 2004; Dilley et al., 2005; D'Imperio et al., 2007; Gili Fivela, 2002; Prieto, 2009; Prieto & Torreira, 2007; Shue et al., 2010; Xu, 1998, 1999, 2001; Xu & Xu, 2005). The alignment has been found to be also sensitive to accent type, vowel length, syllable structure and focus position, as well as being language- or even dialect-dependent (Arvaniti & Garding, 2007; Atterer & Ladd, 2004; D'Imperio et al., 2007; Leyden & Heuven, 2006). The perceptual significance of such alignment is only beginning to be investigated (Dilley, 2005; Niebuhr, 2007; Tokuma & Xu, 2009, 2011). Niebuhr (2007) found that listeners are sensitive not only to turning-point alignment, but also to changes in the shape of F_0 trajectories *due to the alignment change*. This means that information is not necessarily encoded in terms of F_0 peak timing. Consistent with this possibility is the view that the actual encoding units are the underlying pitch targets that are always synchronized with the syllable, and the alignment pattern is the result of interaction between the underlying targets and other factors such as surrounding targets, stress, position of the syllable, focus and duration, etc. (Xu, 2005; Xu & Wang, 2001). Evidence for target-syllable synchronization and target-context interaction can be seen in Dilley (2005), Gao & Xu (2010), Lu and Xu (2006, 2007), Niebuhr (2007) and Xu (1998, 2001). Nevertheless, more research is needed to further resolve this issue by examining the relation between F_0 alignment, articulatory mechanisms and perceptual sensitivity.

4.3 Ecological validity

An issue that faces most, if not all, of experimental studies is the question of ecological validity. That is, how much of what is observed in a controlled experiment is applicable to everyday speech. To address this question, we need first to clarify what is meant by applicability. When made explicit, it could mean one of several things: a) relevance for explaining the prosody of utterances in real life, b) relevance for processing the prosody of real-life utterances in speech recognition, c) relevance for generating naturalistic prosody in speech synthesis, and d) relevance for other practical purposes such as language teaching and speech therapy. For any of these tasks, it would be crucial to have *predictive knowledge* about prosody. So, the real question about ecological validity of a study should be, can it lead to knowledge that is useful for the explanation, recognition, synthesis or other applications 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.¹

¹ Predictive knowledge can be also viewed as the kind that allows us to launch a rocket into the air, knowing that it will reach the moon rather than falling into the ocean, or to give a medicine to a patient, knowing that it will cure rather than kill him. In speech science, predictive knowledge would be

However, also because of the need for systematic controls, individual studies cannot examine all factors at once, and so each is necessarily limited in scope. For example, in a typical lab experiment, subjects are not asked to produce sentences with strong emotions, unless the study is about emotional expressions. Are the findings, then, still valid in cases where emotion is involved? This kind of question should also be empirically answered. There is some initial evidence that functions like focus can be encoded in parallel with emotions (Xu et al., forthcoming). Bruce and Touati (1992) have demonstrated that prosodic patterns found in read speech in Swedish are also found in spontaneous speech, in which emotions are certainly involved (political debate, radio listener call-in conversation). 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 also at establishing predictive knowledge that can be applied in the explanation, recognition, synthesis and other applications of natural prosody.

4.4 Analysis of spontaneous speech

One possible way to improve ecological validity and capture the rich amount of information carried by prosody is to record spontaneous speech in a natural environment. In this strategy, samples of speech are recorded from spontaneous conversation either face to face (Nakajima & Allen, 1993; Swerts, 1992) or over the phone (Campbell, 2001; Cieri & Liberman, 2006), in the classroom (Tseng et al., 2010) or under other naturalistic conditions (Bruce & Touati, 1992; House, 2005; Tseng et al., 2010). One of the most elaborate efforts is the JST CREST ESP Project in which hundreds of hours of spontaneous conversations were recorded, some of which were real-life daily conversations (Campbell, 2000, 2001, 2004). The potential advantage of spontaneous speech over so-called lab speech is that it allows the observation of various phenomena that are lacking in the latter, including discourse markers (Hirschberg & Litman, 1987; Hansson, 1999), interactive structure (Edlund & Heldner, 2005; Oliveira & Freitas, 2008; Swerts & Ostendorf, 1997), turn taking (Caspers, 2003; de Ruiter et al., 2006; Oliveira & Freitas, 2008; Schafer, 1984), repair (van Wijk & Kempen, 1987), filled pauses (Shriberg & Lickley, 1993) and emotion and attitude (Campbell, 2004).

There are various difficulties in studying spontaneous speech, many of which are discussed in detail by Beckman (1997). In particular, its effectiveness depends much on the method of data collection, just as in more controlled studies discussed above. As cautioned by Beckman (1997:19), “the researcher must carefully attend to many aspects of the elicitation paradigm in order to have any luck in getting spontaneous speech that will be useful for the research purpose.” From the perspective of separating functionality from encoding, spontaneous speech presents compounded difficulty. An inherent stumbling block is that, exactly because spontaneous speech is rich in prosodic functionality, each function needs to be properly labeled. Otherwise what is observed for any particular function would easily be confounded by other

the kind that will allow us to build a robot that sounds like a real human, saying all the right things with all the right tone of voice, or a robot that will understand every word we are saying, and every subtlety in our tone of voice, or at least to the extent we are able to do ourselves, or to apply a teaching method to a language learner, knowing that it will work better than chance, and better than any potential “placebo” effect. There is no doubt that we are still a long way away from such predictive knowledge, and most of the research is not directly concerned with practical issues. But the lessons from other scientific disciplines are that predictive knowledge developed in basic research is the key to real advances in technology.

functions. As explained by Campbell (2004:300): "... each utterance must be evaluated separately for such features as the relationships between speaker and hearer (age, sex, familiarity, rank, politeness, etc.), the degree of commitment to the content of the utterance (citing, recalling, revealing, acting, informing, insisting, etc.), the long-term moods and short-term emotions and the attitudinal states of the speaker, the pragmatic force behind the speech act, the voice-quality underlying the utterance (breathy, relaxed, pressed, forced), and so on." If any of these factors is not properly labeled, there is a danger that other factors cannot be properly recognized either. A brief summary in [Table 3](#) (p. 112) demonstrates the diversity of annotations done in a number of studies on spontaneous speech.

A close examination of the studies shown in [Table 3](#) (p. 112) shows that they are mostly focused on discourse-related functions, including turn taking, boundary marking, filled pauses and self-repairs. Also examined are emotions, speaking style and pitch accent types during discourse. Most of these phenomena, with the exception of marking boundaries of lower strengths and certain types of accents, are not easily elicited under laboratory conditions. They are therefore more suitable for this type of studies.

After the collection of the corpus, various analyses need to be conducted. There are almost endless possible ways to analyze a corpus, but the exact method used would depend on the purpose. Similar to the problem of labeling, the potential challenge in the analysis of spontaneous speech is how to overcome confounding of different factors. Various strategies have been developed. Swerts (1997), for example, used listener perception to first rank the boundary strengths, and then examine the differential cues involved in boundaries of different strengths. Nakajima and Allen (1993) examined prosodic events in map task dialogues. Instead of examining them in terms of turn taking, they classified pitch events in terms of topic shift, continuation, elaboration, etc., which seems to correspond to pitch shift of different sizes across the turns. More methodological innovations in future research should further help overcome the intrinsic difficulties with spontaneous speech.

5 Analysis by Modeling

Potentially the most rigorous test of our understanding of prosody, especially in terms of predictive knowledge, is computational modeling, which, though also one type of experimental approach, is discussed separately because its importance has in general not been duly recognized. In modeling, attention to detail can be pushed to the limit, because our knowledge is checked against all the minutiae of reality. To computationally model prosody, quantitative algorithms need to be developed to generate continuous prosodic events whose every detail can be compared to that of real speech. The development of each algorithm is based on a particular understanding of or assumption about an underlying mechanism of prosody. Of course, the rigor of computational modeling as a research tool is not automatically guaranteed, and it is in fact more often not fully realized. This is because its effectiveness depends on a number of critical aspects of the modeling process, including, in particular, (a) description vs. prediction, (b) method of evaluation, and (c) degree of freedom and level of control.

5.1 Descriptive vs. predictive modeling

If the goal of modeling prosody is only to find close mathematical representations of intonation patterns of individual utterances, the most straightforward way is probably to use polynomial functions of various types (simple, spline, piecewise linear, etc.) to fit the F_0 contour of each

utterance. This has been done in a number of studies for tone and intonation (Andruski & Costello, 2004; Chen & Chang, 1992; de Ruiter, 2008; Gandour et al., 1999; Grab et al., 2007; Hirst et al., 2000; Liu et al., 2006). With polynomial fitting, each section of a complex contour can be represented by a set of coefficients, which may drastically reduce the amount of representations from the original point-by-point data. A critical question about polynomial representations, however, is whether they are linguistically meaningful and whether they can be used in predictive modeling, i.e., serving as categorical parameters that can be generalized to other instances of the same category. In the aforementioned studies, the polynomial coefficients are used only in statistic comparisons or classifications, but not in predictive synthesis. So, the predictiveness of models based on polynomial fitting is still unknown. Similarly, a number of other studies that use non-polynomial models also only fitted the F_0 contours of specific utterances without predicting new F_0 contours (Anderson et al., 1984; Bellegarda et al., 2001; Fujisaki, 1983, 1988, 1992; Fujisaki et al., 1994, 2003; Mixdorff & Fujisaki, 1997, 2000; Pierrehumbert, 1981; Taylor, 2000a; van Santen et al., 2005; Veronis et al., 1998). When the modeling experiment is non-predictive, it is hard to know how much of the abstract characteristics of each prosodic category have been captured.

Whether for the sake of building a practical synthesis system or increasing knowledge about prosody, therefore, it is imperative to develop models that are predictive. Predictiveness, however, can be achieved at different levels. The highest degree of predictiveness would be found in 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 developed, of course. The next best would be something akin to a concept-to-speech system (McKeown & Pan 2000; Taylor, 2000b; Young & Fallside, 1979), which, though also very tantalizing, seems to be still far from materialization. Part of the difficulty of concept-to-speech is the translation of concepts into proper functional prosodic categories. But this is also the problem facing systems designed to achieve predictability at an even lower level. That is, 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 system then generate prosodic forms that fit closely to those of the original? From the perspective of analysis by modeling, performing such prediction is actually a process of testing both the validity of the predictors and the assumed encoding mechanisms.

[Table 4](#) (p. 113-114) shows a summary of the predictors (also known as input features) tested in various modeling studies. Note that some of the studies in the table have used a large number of predictors, e.g., Bellegarda et al. (2001), Mixdorff and Jokisch (2001), Mohler and Conkie (1998) and Sun (2002). In these cases, for the sake of developing predictive knowledge, it would be desirable to not only assess the overall quality of data fitting, but also examine the contribution of each predictor. Note also that the performance of a particular predictor does not necessarily reflect its importance, because the performance should also be closely related to its modeling implementation based on the assumed encoding strategy. As a tool for hypothesis testing, prosodic modeling therefore should be a continuous process of not only trying to identify all the relevant predictors, but also searching for the proper encoding mechanisms, which are not necessarily shared across the predictors (cf. Xu, 2005 for some hypotheses).

5.2 Method of evaluation

Whether a model is descriptive or predictive, its effectiveness needs to be evaluated one way or another. The methods of evaluation have been quite diverse, ranging from informal visual inspection and listening to various formal objective and subjective evaluations. The goal of

objective evaluations is to measure goodness of fit between synthetic and original prosodic forms. There have been three major methods: Root Mean Square Error (RMSE) (Fujisaki et al., 2005; Kochanski et al., 2003; Ni et al. 2006; Prom-on et al., 2009; Raidt et al., 2004; Sakurai et al., 2003; Sun, 2006) and Pearson's correlation or correlation for short (Mixdorff & Jokisch 2001; Prom-on et al., 2009; Raidt et al., 2004; Sun, 2006), and mean absolute frequency deviation across each utterance (Bellegarda et al., 2001). Among the three, RMSE and correlation are the most used. Hermes (1998) has shown that both measures are effective. For RMSE, the unit of measurement varies from Hz (Mohler & Conkie, 1998; Morlect et al., 2001; Sun, 2002) to semitone (Prom-on et al., 2009; Raidt et al., 2004) to $\ln(F_0)$ (Fujisaki et al., 2005; Gu et al., 2006; Ni et al., 2006; Sakurai et al., 2003), which makes cross-study comparison difficult. Although $\ln(F_0)$ and semitones can be mutually converted from one to the other because both are logarithmic ($\text{semitone}_{\text{RMSE}} = 12 \ln(F_0)_{\text{RMSE}} / \ln(2)$), RMSE values in Hz cannot be properly converted to a logarithmic scale, because of lack of reference, see equation 1. The advantage of logarithmic scale for intonation has been shown in terms of both production (Fujisaki, 2003; Nolan, 2003; Xu & Sun, 2002) and perception (Traunmüller & Eriksson, 1994). Given such logarithmic nature, RMSE values in Hz need to be interpreted differently depending on the average pitch of the speaker. For example, an RMSE value of 10 Hz (≈ 1.39 st) for a male speaker with a mean F_0 of $F_{0\text{reference}} = 120$ Hz is roughly equivalent to that of 18 Hz (≈ 1.36 st) for a female voice with a mean F_0 of $F_{0\text{reference}} = 220$ Hz.

$$\text{semitone} = 12 \log_2(F_0 / F_{0\text{reference}}) \quad (1)$$

In addition to the objective evaluations, the effectiveness of a model can also be perceptually evaluated. The perceptual evaluations fall into two major types. In the first type, the naturalness of the synthetic prosody is evaluated. Various methods have been used, including direct judgment of naturalness (Sun, 2002), judgment of whether the prosody is synthetic or natural (Prom-on et al., 2009), and similarity rating (Ni et al., 2006). Some studies (e.g., Mohler & Conkie, 1998) have made use of mean opinion score (MOS) a measurement designed for evaluating audio media quality, which employs a 5-level scale:

Mean opinion score (MOS)		
MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

The MOS is then calculated as the arithmetic mean of all the individual scores. This method seems to have the potential of standardizing formal naturalness evaluations. But its advantage has yet to be fully demonstrated.

In the second type of evaluation, what is assessed is the perceptual accuracy of the hypothetical categories being modeled, such as lexical tone, lexical stress, pitch accent, focus, boundary strength, sentence modality, type of emotion and attitude, etc. Surprisingly, only two of the studies reviewed here performed this kind of evaluations (Morlec et al., 2001; Prom-on et al., 2009). All the other studies that performed perceptual evaluations only assessed perceptual naturalness of synthetic prosody. For the sake of improving the predictiveness of the models and

using them as effective hypothesis testing tools, it is highly desirable in future research to include perceptual identification of the input categories as one of the evaluation methods.

5.3 Degrees of freedom and Level of control

Degrees of freedom (df) refers to the number of free parameters needed to completely specify a model. These are the parameters that must be estimated in order for the model to generate the intended output. Degrees of freedom is seldom discussed, however, and in many theories and models the number of free parameters is kept implicit. The importance of degrees of freedom is that it reflects assumptions about issues such as what is speaker-controlled and what is articulatorily mandatory, what is functionally encoded and what is perceptually relevant. Making these assumptions explicit will allow various tradeoffs related to degrees of freedom to be directly evaluated. The following are some common issues closely related to degrees of freedom.

- 1) What are the assumed basic prosodic events and how are they specified? For example, in an increasing order of df, the assumed basic events can be single points, straight lines, piece-wise linear shapes or complex contours.
- 2) Are the properties of the basic events fixed or variable? Fixed properties often entail lower df than variable properties. For variable properties, how many parameters are needed to specify the variability?
- 3) How is the timing of the basic events specified? The df would be larger if the timing is freely variable and thus needs to be estimated than if the timing is fixed relative to segmental events such as the onset, offset or the entire interval of vowel, rhyme, syllable or word.
- 4) Are transitions needed between the basic events? If yes, how many parameters are needed to specify each transition?
- 5) Are global parameters, such as top and base lines, needed to specify each local contour?

[Table 5](#) (p. 115) shows a comparison of degrees of freedom in a number of models for controlling local F_0 contours. What is shown here, however, is only degrees of freedom for generating local pitch events. Note that even at this level, degrees of freedom should not be used as the sole criterion for judging a model, and it is not the case that smaller degrees of freedom is always better. What is more crucial is whether each degree of freedom has sufficient justifications. Ideally the degrees of freedom should reflect the assumed human control, at least at the level that the model tries to simulate. For example, if transitions between basic events are specified by one or more free parameters, it would be helpful to explicate the articulatory nature of such controlled transitions. Furthermore, degrees of freedom also has to do with the level of control the model tries to simulate, e.g., at the level of the muscle (Fujisaki et al., 2005), the underlying target and its articulatory approximation (Kochanski & Shih, 2003; Prom-on et al., 2009), the decomposed holistic contours (Baily et al., 2005) or surface contours (Pierrehumber, 1981; Taylor 2000; van Santen et al. 2005).

For more global events related to higher functions as shown in [Table 1](#) (p. 110-111), more degrees of freedom is certainly needed. For each contrastive prosodic component above the basic units, at least a single degree of freedom is required to simulate its coding. Not implementing this degree of freedom would result in a decrease in the functionality of the model.

6 Concluding remarks: Need for linking and integration

We have seen through this brief review that much effort has been made to build up a body of predictive knowledge on speech prosody. The goal of this review has been to highlight the differences between the various methodological approaches in terms of their effectiveness in solving the lack of reference problem. We have seen that there is mounting evidence that focus, boundary marking and modality are three highly likely functions, and there seems to be converging evidence for their distinctive prosodic coding. The picture for other functions is less clear. At the same time, we can also see a striking disconnect between subareas of prosody research. For example, although I have used the term analysis by modeling, the truth is that most of modeling activities are not conducted with the aim to test any existing theory or to develop a new theory, but are rather done for modeling's own sake. As a result, major theories of prosody have yet to be put through the most rigorously test, namely, being forced to make predictions about all the prosodic details that can be checked against real speech, or to recognize prosodic functions from real speech. Similar disconnect, though to a lesser degree, can be also seen between perception-oriented efforts and production-oriented ones. It is therefore highly desirable for there to be much more linking between the subareas and different approaches and hopefully also true integrations. If such linking and integration are to occur, the next ten years should see some truly accelerated advances in both theoretical development and practical applications.

Acknowledgement

I would like to thank Plinio Barbosa, Paul Gettel and Santitham Prom-on for their comments on an earlier version of the paper. All errors are mine, of course.

REFERENCES

- Anderson M, Pierrehumbert J and Liberman M. Synthesis by rule of English intonation patterns In: Proceedings of Proceedings of ICASSP, San Diego, CA, 1984. 77-80.
- Andruski J and Costello J. Using polynomial equations to model pitch contour shape in lexical tones: An example from Green Mong. *Journal of the International Phonetic Association* 2004;34:125-140.
- Anttila A, Adams M and Speriosu M. The role of prosody in the English dative alternation. *Language and Cognitive Processes* 2010;25:946-981.
- Arvaniti A and Garding G. Dialectal variation in the rising accents of American English. In: J Cole and J Hualde, editors. *Laboratory Phonology 9*, The Hague: Mouton de Gruyter, 2007. 547-576.
- Atterer M and Ladd DR. On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German. *Journal of Phonetics* 2004;32:177-197.
- Auberge V and Cathiard M. Can we hear the prosody of smile. *Speech Communication* 2003;40:87-97.
- Avesani C and Vayra M. Broad, narrow and contrastive focus in Florentine Italian In: Proceedings of The 15th International Congress of Phonetic Sciences, Barcelona, 2003. 1803-1806.
- Bagshaw PC. An investigation of acoustic events related to sentential stress and pitch accents, in English. *Speech Communication* 1993;13:333-342.
- Bailly G and Holm B. SFC: a trainable prosodic model. *Speech Communication* 2005;46:348-364.
- Bänziger T and Scherer KR. The role of intonation in emotional expressions. *Speech Communication* 2005;46:252-267.

- Barbosa PA. From syntax to acoustic duration: A dynamical model of speech rhythm production. *Speech Communication* 2007;49:725-742.
- Barbosa PA. Detecting changes in speech expressiveness in participants of a radio program In: *Proceedings of Interspeech 2009*, Brighton, UK, 2009. 2155-2158.
- Barbosa PA. and Bailly G. Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication* 1994;15 (1-2):127-137.
- Baum KM and Nowicki S. Perception of emotion: Measuring decoding accuracy of adult prosodic cues varying in intensity. *Journal of Nonverbal Behavior* 1998;22:89-107.
- Baumann S, Becker J, Grice M and Mücke D. Tonal and articulatory marking of focus in German In: *Proceedings of The 16th International Congress of Phonetic Sciences*, Saarbrücken, 2007. 1029-1032.
- Beckman ME and Edwards J. Lengthenings and shortenings and the nature of prosodic constituency. In: J Kingston and ME Beckman, editors. *Papers in Laboratory Phonology 1 — Between the Grammar and Physics of Speech*, Cambridge: Cambridge University Press, 1990. 152-178.
- Beckman ME. A typology of spontaneous speech. In: Y Sagisaka, N Campbell and N Higuchi, editors. *Computing Prosody: Computational Models for Processing Spontaneous Speech*, New York: Springer Verlag, 1997. 7-26.
- Bellegarda J, Silverman K, Lenzo K and Anderson V. Statistical prosodic modeling: from corpus design to parameter estimation. *IEEE Transactions on Speech Audio Process* 2001;9:52-66.
- Beller G, Obin N and rodet X. Articulation Degree as a Prosodic Dimension of Expressive Speech In: *Proceedings of Speech Prosody 2008*, 2008.
- Bentin S, Ram F and Leonard K. Chapter 11 Phonological Awareness, Reading, and Reading Acquisition: A Survey and Appraisal of Current Knowledge. In: editors. *Advances in Psychology*. Volume 94: North-Holland, 1992. 193-210.
- Berkovits R. Utterance-final lengthening and the duration of final-stop closures. *Journal of Phonetics* 1993;21:479-489.
- Berkovits R. Durational Effects in Final Lengthening, Gapping, and Contrastive Stress. *language and speech* 1994;37:237-250.
- Beysade C, Hemforth B, Marandin J-M and Portes C. Prosodic markings of information focus in French In: *Proceedings of Interface Discours & Prosodie 2009*, Paris, 2009. 109-122.
- Birch S and Clifton C. Focus, accent, and argument structure: effects on language comprehension. *Language and Speech* 1995;38:365-391.
- Bock J and Mazzella J. Intonational marking of given and new information: Some consequences for comprehension. *Memory & Cognition* 1983;11:64-76.
- Bolinger DL. *Forms of English: Accent, Morpheme, Order*, Cambridge, Massachusetts: Harvard University Press; 1965.
- Bolinger D. *Intonation and its parts: melody in spoken English*, Palo Alto: Stanford University Press; 1986.
- Bolinger D. *Intonation and Its Uses -- Melody in Grammar and Discourse*, Stanford, California: Stanford University Press; 1989.
- Botinis A, Bannert R and Tatham M. Contrastive tonal analysis of focus perception in Greek and Swedish. In: A Botinis, editors. *Intonation : analysis, modelling and technology*, Boston: Kluwer Academic Publishers, 2000. 97-116.
- Botinis A, Fourakis M and Gawronska B. Focus identification in English, Greek and Swedish In: *Proceedings of The 14th International Congress of Phonetic Sciences*, San Francisco, 1999. 1557-1560.
- Botinis A, Granström B and Möbius B. Developments and paradigms in intonation research. *Speech Communication* 2001;33:263-296.
- Braun B and Tagliapietra L. The role of contrastive intonation contours in the retrieval of contextual alternatives. *Language and Cognitive Processes* 2010;25:1024-1043.

- Brazil DM, Coulthard M and Johns C. *Discourse Intonation and Language Teaching*, London: Longman; 1980.
- Breitenstein C, Van Lancker D and Daum I. The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition & Emotion* 2001;15:57-79.
- Bulut M and Narayanan S. On the robustness of overall F0-only modifications to the perception of emotions in speech. *Journal of the Acoustical Society of America* 2008;123:4547-4558.
- Bruce G and Touati P. On the analysis of prosody in spontaneous speech with exemplification from Swedish and French. *Speech Communication* 1992;11:453-458.
- Bruce G. Developing the Swedish intonation model. Lund University, Dept. of Linguistics Working Papers 1982a;22:51-116.
- Bruce G. Textual aspects of prosody in Swedish. *Phonetica* 1982b;39:274-287.
- Büring D. On D-Trees, Beans, and B-Accents. *Linguistics and Philosophy* 2003;26:511-545.
- Büring D. Focus Projection and Default Prominence. In: V Molnar and S Winkler, editors. *The Architecture of Focus*, 2006.
- Büring D. Semantics, Intonation and Information Structure. In: G Ramchand and C Reiss, editors. *The Oxford Handbook of Linguistic Interfaces*: Oxford University Press, 2007.
- Byrd D and Saltzman E. Intra-gestural dynamics of multiple prosodic boundaries. *Journal of Phonetics* 1998;26:173-199.
- Byrd D and Saltzman E. The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics* 2003;31:149-180.
- Calhoun S. How does informativeness affect prosodic prominence? *Language and Cognitive Processes* 25:1099-1140.
- Campbell N. Automatic detection of prosodic boundaries in speech. *Speech Communication* 1993;13:343-354.
- Campbell WN and Marumoto T. Automatic labelling of voice quality in speech databases for synthesis In: *Proceedings of ICSLP-2000, Beijing, 2000*. 468-471.
- Campbell N. Building a Corpus of Natural Speech - and Tools for the Processing of Expressive Speech - the JST CREST ESP Project In: *Proceedings of Eurospeech 2001, 2001*. 1525-1528.
- Campbell N. Databases of expressive speech. *Journal of Chinese Language and Computing* 2004;14:295-304.
- Carlson, K., Clifton, C. and Frazier, L. (2001). Prosodic Boundaries in Adjunct Attachment. *Journal of Memory and Language* 45(1): 58-81.
- Caspers J. Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics* 2003;31:251-276.
- Chaffe W. Language and consciousness. *Language* 1974;50:111-133.
- Chaffe W. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In: C Li, editors. *Subject and Topic*, New York: Academic Press, 1976. 25-55.
- Chahal D. Phonetic Cues to Prominence in Lebanese Arabic In: *Proceedings of The 15th International Congress of Phonetic Sciences, Barcelona, 2003*. 2067-2070.
- Chen S-H and Chang S. A statistical model based fundamental frequency synthesizer for Mandarin speech. *Journal of the Acoustical Society of America* 1992;92:114-120.
- Chen A and Destruel E. Intonational encoding of focus in Toulousian French In: *Proceedings of Speech Prosody 2010, Chicago, 2010*.
- Chen A, Gussenhoven C and Rietveld T. Language-specificity in the perception of paralinguistic intonational meaning. *Language and Speech* 2004;47:311-349.
- Chen S-w, Wang B and Xu Y. Closely related languages, different ways of realizing focus In: *Proceedings of Interspeech 2009, Brighton, UK, 2009*. 1007-1010.

- Chen Y. Durational Adjustment under Contrastive Focus in Standard Chinese. *Journal of Phonetics* 2006;34:176-201.
- Cho T. Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics* 2004;32:141-176.
- Cho T and McQueen JM. Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics* 2005;33:121-157.
- Christophe A, Gout A, Peperkamp S and Morgan J. Discovering words in the continuous speech stream: the role of prosody. *Journal of Phonetics* 2003;31:585-598.
- Chuenwattanapranithi S, Xu Y, Thipakorn B and Maneewongvatana S. Encoding emotions in speech with the size code — A perceptual investigation. *Phonetica* 2008;65:210-230.
- Cieri C and Liberman M. More Data and Tools for More Languages and Research Areas: A Progress Report on LDC Activities. LREC 2006: Fifth International Conference on Language Resources and Evaluation, 2006.
- Cole J, Mo Y and Baek S. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes* 2010;25:1141-1177.
- Cooper WE, Eady SJ and Mueller PR. Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America* 1985;77:2142-2156.
- Cosmides L. Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception & Performance* 1983;9:864-881.
- Cruttenden A. *Intonation* (second edition): Cambridge University Press; 1997.
- Crystal D. *Prosodic Systems and Intonation in English*, London: Cambridge University Press; 1969.
- Cutler A, Dahan D and van Donselaar W. Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 1997;40:141-201.
- Dankovicova J, House J, Crooks A and Jones K. The Relationship between Musical Skills, Music Training, and Intonation Analysis Skills. *Language & Speech* 2007;50:177-225.
- de Jong KJ. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America* 1995;97:491-504.
- de Jong K. Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *Journal of Phonetics* 2004;32:493-516.
- de Jong K and Zawaydeh B. Comparing stress, lexical focus, and segmental focus: patterns of variation in Arabic vowel duration. *Journal of Phonetics* 2002;30:53-75.
- de Ruiter JP, Mitterer H and Enfield NJ. Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language* 2006;82:515-535.
- de Ruiter LE. How useful are polynomials for analyzing intonation? In: *Proceedings of Interspeech 2008, Brisbane, 2008*. 785-789.
- Delais-Roussarie E, Rialland A, Doetjes J and Marandin JM. The prosody of post-focus sequences in French In: *Proceedings of The 1st International Conference on Speech Prosody, Aix-en-Provence, France, 2002*. 239-242.
- Delattre P, Poenack E and Olsen C. Some characteristics of German intonation for the expression of continuation and finality. *Phonetica* 1965;13:134-161.
- Di Cristo A and Jankowski J. Prosodic organisation and phrasing after focus in French In: *Proceedings of The 14th International Congress of Phonetic Sciences, San Francisco, 1999*. 1565-1568.
- Dilley LC, Ladd DR and Schepman A. Alignment of L and H in bitonal pitch accents: testing two hypotheses. *Journal of Phonetics* 2005;33:115-119.
- D'Imperio M. Focus and tonal structure in Neapolitan Italian. *Speech Communication* 2001;33:339-356.
- D'Imperio M, Espesser R, Løevenbrück H, Menezes C, Nguyen N and Welby P. Are tones aligned with articulatory events? Evidence from Italian and French. In: J Cole and J Hualde, editors. *Papers in Laboratory Phonology IX: Change in Phonology, The Hague: Mouton de Gruyter, 2007*. 577-608.

- Dohen M and Lævenbruck H. Pre-focal rephrasing, focal enhancement and post-focal deaccentuation in French In: Proceedings of The 8th International Conference on Spoken Language Processing, Jeju, Korea, 2004. 1313-1316.
- Donati C and Nespors M. From focus to syntax. *Lingua* 2003;113:1119-1142.
- Dorn A and Ní Chasaide A. Effects of focus on f0 and duration in Irish (Gaelic) declaratives. In: Proceedings of Interspeech 2011, Florence, 2011.
- Downing LJ. Focus and prominence in Chichewa, Chitumbuka and Durban Zulu. *ZAS Papers in Linguistics (ZASPiL)* 2008;49:47-65.
- Eady SJ and Cooper WE. Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America* 1986;80:402-416.
- Eady SJ, Cooper WE, Klouda GV, Mueller PR and Lotts DW. Acoustic characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and Speech* 1986;29:233-251.
- Edlund J and Heldner M. Exploring Prosody in Interaction Control. *Phonetica* 2005;62:215-226.
- Edwards JR, Beckman ME and Fletcher J. The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America* 1991;89:369-382.
- Face TL. Narrow Focus intonation in Castilian Spanish absolute interrogatives. *Journal of Language and Linguistics* 2006;5:295-311.
- Ferriera F. The creation of prosody during sentence production. *Psychological Review* 1993;100:233-253.
- Féry C and Kügler F. Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics* 2008;36:680-703.
- Fery C, Skopeteas S and Hornig R. Cross-linguistic comparison of prosody, syntax and information structure in a production experiment on localising expressions. *Transactions of the Philological Society* 2010;108:329-351.
- Fougeron C. Articulatory properties of initial segments in several prosodic constituents in French. *Journal of phonetics* 2001;29:109-135.
- Frick RW. Communicating emotion: The role of prosodic features. *Psychological Bulletin* 1985;97:412-429.
- Fujisaki H. Dynamic characteristics of voice fundamental frequency in speech and singing. In: PF MacNeilage, editors. *The Production of Speech*, New York: Springer-Verlag, 1983. 39-55.
- Fujisaki H. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In: O Fujimura, editors. *Vocal Physiology: Voice Production*, New York: Raven Press, Ltd., 1988. 347-355.
- Fujisaki H. Modeling the process of fundamental frequency contour generation. In: Y Tohkura, E Vatikiotis-Bateson and Y Sagisaka, editors. *Speech Perception, Production and Linguistic Structure*, Amsterdam: IOS Press, 1992. 313-326.
- Fujisaki H. Prosody, Information, and Modeling — with Emphasis on Tonal Features of Speech In: Proceedings of Workshop on Spoken Language Processing, 2003. 5-14.
- Fujisaki H, Ohno S, Nakamura K, Fuirao M and Gurlekian J. Analysis of accent and intonation in Spanish based on a quantitative model In: Proceedings of International Conference on Spoken Language Processing, Yokohama, 1994. 355-358.
- Fujisaki H, Ohno S, Osame M, Sakata M and Hirose K. Prosodic characteristics of a spoken dialogue for information query In: Proceedings of International Conference on Spoken Language Processing, Yokohama, 1994. 1103-1106.
- Fujisaki H, Wang C, Ohno S and Gu W. Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model. *Speech communication* 2005;47:59-70.
- Gao H and Xu Y. Ambisyllabicity in English: How real is it? In: Proceedings of The 9th Phonetics Conference of China (PCC2010), Tianjin, 2010.
- Gårding E. Sentence intonation in Swedish. *Phonetica* 1979;36:207-215.

- Gårding E. Speech act and tonal pattern in Standard Chinese. *Phonetica* 1987;44:13-29.
- Gerard C and Dahan D. Durational variations in speech and didactic accent during reading. *Speech Communication* 1995;16:293-311.
- Gili Fivela B. Tonal alignment in two Pisa Italian peak accents In: Proceedings of The 1st International Conference on Speech Prosody, Aix-en-Provence, France, 2002. 339-342.
- Gobl C and Chasaide AN. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 2003;40:189-212.
- Grabe E, Kochanski G and Coleman J. Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech* 2007;50:281-310.
- Grice M, Baumann S and Jagdfeld N. Tonal association and derived nuclear accents--The case of downstepping contours in German. *Lingua* 2009;119:881-905.
- Grønnum N. A Danish phonetically annotated spontaneous speech corpus (DanPASS). *Speech Communication* 2009;51:594-603.
- Gu W, Hirose K and Fujisaki H. Modeling the effects of emphasis and question on fundamental frequency contours of Cantonese utterances. *Audio, Speech, and Language Processing, IEEE Transactions on* 2006;14:1155-1170.
- Gu W and Lee T. Effects of tonal context and focus on Cantonese F0 In: Proceedings of The 16th International Congress of Phonetic Sciences, Saarbrücken, 2007. 1033-1036.
- Gussenhoven C. Types of focus in English. In: C Lee, M Gordon and D Büring, editors. *Topic and Focus: Cross-linguistic Perspectives on Meaning and Intonation*, New York: Springer, 2007. 83-100.
- Gussenhoven C, Repp BH, Rietveld A, Rump HH and Terken J. The perceptual prominence of fundamental frequency peaks. *Journal of the Acoustical Society of America* 1997;102:3009-3022.
- Gussenhoven C and Rietveld ACM. Intonation contours, prosodic structure and preboundary lengthening. *Journal of Phonetics* 1992;20:283-303.
- Halliday MAK. *Intonation and Grammar in British English*, The Hague: Mouton; 1967.
- Hansson P. Prosodic correlates of discourse markers in dialogue In: Proceedings of DIAPRO-1999, 1999. 99-104.
- Hanssen J, Peters J and Gussenhoven C. Prosodic Effects of Focus in Dutch Declaratives In: Proceedings of Speech Prosody 2008, Campinas, Brazil, 2008. 609-612.
- Hartmann K. Focus and tone. *Acta Linguistica Hungarica* 2008;55:415-426.
- Hartmann K and Zimmermann M. Focus Strategies in Chadic: The Case of Tangale Revisited. *Studia Linguistica* 2007;61:95-129.
- Heldner M and Strangert E. Temporal effects of focus in Swedish. *Journal of Phonetics* 2001;29:329-361.
- Heldner M. On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics* 2003;31:39-62.
- Hellmuth S. Focus-related pitch range manipulation (and peak alignment effects) in Egyptian Arabic In: Proceedings of Speech Prosody 2006, Dresden, Germany, 2006. PS4-12-164.
- Hermes DJ. Measuring the Perceptual Similarity of Pitch Contours. *J Speech Lang Hear Res* 1998;41:73-82.
- Hermes DJ and Rump HH. Perception of prominence in speech intonation induced by rising and falling pitch movements. *Journal of the Acoustical Society of America* 1994;96:83-92.
- Hirschberg J and Litman D. Now let's talk about now: Identifying cue phrases intonationally In: Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, 1987. 163-171.
- Hirst DJ, Di Cristo A and Espesser R. Levels of representation and levels of analysis for intonation. In: M Horne, editors. *Prosody: Theory and Experiment*, Dordrecht: Kluwer, 2000. 51-87.

- Hitzeman J, Black AW, Mellish C, Oberlander J, Poesio M and Taylor P. An Annotation Scheme for Concept-to-Speech Synthesis In: Proceedings of European Workshop on Natural Language Generation, Toulouse, France, 1999. 59-66.
- Ho AT. Intonation variation in a Mandarin sentence for three expressions: interrogative, exclamatory and declarative. *Phonetica* 1977;34:446-457.
- Ho AT. Mandarin tones in relation to sentence intonation and grammatical structure. *Journal of Chinese Linguistics* 1976;4:1-13.
- Hockett CF. A course in modern linguistics, New York: MacMillan; 1958.
- House D. Phrase-final rises as a prosodic feature in wh-questions in Swedish human-machine dialogue. *Speech Communication* 2005;46:268-283.
- Ipek C. 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, 2011.
- Ishihara S. Intonation and interface conditions. Massachusetts Institute of Technology; 2003.
- Jackendoff R. Semantics in Generative Grammar, Cambridge, MA: MIT Press; 1972.
- Jannedy S. Prosodic Focus in Vietnamese. In: S Ishihara, S Jannedy and A Schwarz, editors. *Interdisciplinary Studies on Information Structure (ISIS)*: Potsdam University Press, 2007. 209-230.
- Jin S. An Acoustic Study of Sentence Stress in Mandarin Chinese [Ph.D. dissertation]. The Ohio State University; 1996.
- Kabagema-Bilan E, LÚpez-JimÈnez B, Truckenbrodt H, University Tb and Berlin ZAS. Multiple focus in Mandarin Chinese. *Lingua* 2011;In Press, Corrected Proof
- Kingdon R. The groundwork of English intonation, London: Longman; 1958.
- Klatt DH. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics* 1975;3:129-140.
- Kjelgaard MM and Speer SR. Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language* 1999;40:153-194.
- Kochanski G, Grabe E, Coleman J and Rosner B. Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America* 2005;118:1038-1054.
- Kochanski G and Shih C. Prosody modeling with soft templates. *Speech Communication* 2003;39:311-352.
- Kochanski G, Shih C and Jing H. Quantitative measurement of prosodic strength in Mandarin. *Speech Communication* 2003;41:625-645.
- Kohler KJ. Modelling prosody in spontaneous speech. In: Y Sagisaka, N Campbell and N Higuchi, editors. *Computing Prosody*, New York: Springer, 1997. 187-210.
- Kiss KÉ. Topic and focus: Two structural positions associated with logical functions in the left periphery of the Hungarian sentence. *Acta Linguistica Hungarica* 2008;55:287-296.
- Krifka M. Basic notions of information structure. *Acta Linguistica Hungarica* 2008;55:243-276.
- Krivokapic J. Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics* 2007;35:162-179.
- Kügler F and Skopeteas S. 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, 2007.
- Ladd DR. Declination 'reset' and the hierarchical organization of utterances. *Journal of the Acoustical Society of America* 1988;84:530-544.
- Ladd DR, Silverman KEA, Tolkmitt F, Bergmann G and Scherer KR. Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *Journal of the Acoustical Society of America* 1985;78:435-444.
- Ladd DR, Verhoeven J and Jacobs K. Influence of adjacent pitch accents on each other's perceived prominence: two contradictory effects. *Journal of Phonetics* 1994;22:87-99.

- Lecumberri MLG. Perceptibility of nuclear focus in English. *Barcelona English language and literature studies* 1997:175-186.
- Lee Y-c and Xu Y. Phonetic Realization of Contrastive Focus in Korean In: *Proceedings of Speech Prosody 2010*, Chicago, 2010. 100033:1-4.
- Lehiste I. The phonetic structure of paragraphs. In: A Cohen and SEG Nootboom, editors. *Structure and process in speech perception*, New York: Springer-Verlag, 1975. 195-206.
- Lehiste I, Olive JP and Streeter LA. Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America* 1976;60:1199-1202.
- Leyden K and van Heuven VJ. On the Prosody of Orkney and Shetland Dialects. *Phonetica* 2006;63:149-174.
- Lieberman IY, Shankweiler D and Liberman AM. The Alphabetic Principle and Learning to Read. In: D Shankweiler and IY Liberman, editors. *Phonology and Reading Disability: Solving the Reading Puzzle*, Ann Arbor: University of Michigan Press, 1990.
- Lin M. On production and perception of boundary tone in Chinese intonation In: *Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing, 2004. 125-129.
- Liu F, Surendran D and Xu Y. Classification of statement and question intonations in Mandarin In: *Proceedings of Speech Prosody 2006*, Dresden, Germany, 2006. PS5-25_0232.
- Liu F and Xu Y. Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 2005;62:70-87.
- Liu F and Xu Y. Question intonation as affected by word stress and focus in English In: *Proceedings of The 16th International Congress of Phonetic Sciences*, Saarbrücken, 2007. 1189-1192.
- Liu F and Xu Y. The Neutral Tone in Question Intonation in Mandarin In: *Proceedings of Interspeech 2007*, Antwerp, 2007. 630-633.
- Mann VA and Brady S. Reading Disability: The Role of Language Deficiencies. *Journal of Consulting and Clinical Psychology* 1988;56:811-816.
- Makarova V. Perceptual correlates of sentence-type intonation in Russian and Japanese. *Journal of Phonetics* 2001;29:137-154.
- Mattingly I. Reading, the linguistic process, and linguistic awareness. In: J Kavanagh and I Mattingly, editors. *Language by ear and by eye: The relationships between speech and reading*, Cambridge, MA: MIT Press, 1972. 133-147.
- Mady K and Kleber F. Variation of pitch accent patterns in Hungarian In: *Proceedings of Speech Prosody 2010*, Chicago, 2010.
- Mauss IB and Robinson MD. Measures of emotion: A review. *Cognition & Emotion* 2009;23:209-237.
- McKeown KR and Pan S. Prosody modelling in concept-to-speech generation: methodological issues. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 2000;358:1419-1431.
- Metusalem R and Ito K. The role of L+H* pitch accent in discourse construction In: *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 2008.
- McRoberts GW, Studdert-Kennedy M and Shankweiler DP. The role of fundamental frequency in signalling linguistic stress and affect: Evidence for a dissociation. *Perception & Psychophysics* 1995;57:159-174.
- Mixdorff H. A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters In: *Proceedings of ICASSP 2000*, Istanbul, Turkey, 2000. 1281-1284.
- Mixdorff H and Fujisaki H. Automated quantitative analysis of F0 contours of utterances from a German ToBI-labeled speech database In: *Proceedings of The '97 Eurospeech*, Rhodes, Greece, 1997. 187-190.
- Mixdorff H and Jokisch O. Building An Integrated Prosodic Model of German In: *Proceedings of Eurospeech 2001*, 2001.

- Mohler G and Conkie A. Parametric modeling of intonation using vector quantization In: Proceedings of The Third ESCA/COCOSDA Workshop on Speech Synthesis, Blue Mountains, NSW, Australia, 1998.
- Morlec Y, Bailly G and Aubergé V. Generating prosodic attitudes in French: Data, model and evaluation. *Speech Communication* 2001;33:357-371.
- Morton EW. On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist* 1977;111:855-869.
- Most RB and Saltz E. Information structure in sentences: New information. *Language & Speech* 1979;22:89-95.
- Mozziconacci SJL. Modeling emotion and attitude in speech by means of perceptually based parameter values. *User Modeling and User-Adapted Interaction* 2001;11:297-326.
- Murray IR and Arnott JL. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America* 1993;93:1097-1108.
- Nakajima S and Allen JF. A study on prosody and discourse structure in cooperative dialogues. *Phonetica* 1993;50:197-210.
- Nakatani LH, O'Connor KD and Aston CH. Prosodic aspects of American English speech rhythm. *Phonetica* 1981;38:84-106.
- Ni J, Kawai H and Hirose K. Constrained tone transformation technique for separation and combination of Mandarin tone and intonation. *Journal of the Acoustical Society of America* 2006;119:1764-1782.
- Niebuhr O. The Signalling of German Rising-Falling Intonation Categories – The Interplay of Synchronization, Shape, and Height. *Phonetica* 2007;64:174-193.
- Nolan F. Intonational equivalence: an experimental evaluation of pitch scales In: Proceedings of The 15th International Congress of Phonetic Sciences, Barcelona, 2003. 771-774.
- Nooteboom SG and Kruyt JG. Accents, focus distribution, and the perceived distribution of given and new information: An experiment. *Journal of the Acoustical Society of America* 1987;82:1512-1524.
- O'Connor JD and Arnold GF. *Intonation of Colloquial English*, London: Longmans; 1961.
- Ofuka E, McKeown JD, Waterman MG and Roach PJ. Prosodic cues for rated politeness in Japanese speech. *Speech Communication* 2000;32:199-217.
- Ogden R and Routarinne S. The Communicative Functions of Final Rises in Finnish Intonation. *Phonetica* 2005;62:160-175.
- Ohala JJ. An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* 1984;41:1-16.
- Ohala JJ. Ethological theory and the expression of emotion in the voice In: Proceedings of ICSLP96, 1996. 1812-1815.
- Oliveira M and Freitas T. Intonation as a cue to turn management in telephone and face-to-face interactions In: Proceedings of Speech Prosody 2008, Campiñas, 2008. 485-488.
- O'Shaughnessy D and Allen J. Linguistic modality effects on fundamental frequency in speech. *Journal of the Acoustical Society of America* 1983;74:1155-1171.
- Ouden Hd, Noordman L and Terken J. Prosodic realizations of global and local structure and rhetorical relations in read aloud news reports. *Speech Communication* 2009;51:116-129.
- Palmer HE. *English Intonation, with Systematic Exercises*, Cambridge: Heffer; 1922.
- Pan H. Focus and Taiwanese unchecked tones. In: C Lee, M Gordon and D Büring, editors. *Topic and Focus: Cross-linguistic Perspectives on Meaning and Intonation*: Springer, 2007. 195-213.
- Patil U, Kentner G, Gollrad A, Kügler F, Féry C and Vasishth S. Focus, word order and intonation in Hindi. *Journal of South Asian Linguistics* 2008;1:55-72.
- Pell MD. Influence of emotion and focus on prosody in matched statements and questions. *Journal of the Acoustical Society of America* 2001;109:1668-1680.

- Pierrehumbert J and Hirschberg J. The meaning of intonational contours in the interpretation of discourse. In: PR Cohen, J Morgan and ME Pollack, editors. *Intentions in Communication*, Cambridge, Massachusetts: MIT Press, 1990. 271-311.
- Pierrehumbert J. Synthesizing intonation. *Journal of the Acoustical Society of America* 1981;70:985-995.
- Pierrehumbert J. *The Phonology and Phonetics of English Intonation* [Ph.D. dissertation]. MIT, Cambridge, MA. [Published in 1987 by Indiana University Linguistics Club, Bloomington]; 1980.
- Pike KL. *The Intonation of American English*, Ann Arbor: University of Michigan Press; 1945.
- Popper K. *The Logic of Scientific Discovery* (translation of *Logik der Forschung*), London: Hutchinson; 1959.
- Potisuk S, Gandour J and Harper MP. Contextual variations in trisyllabic sequences of Thai tones. *Phonetica* 1996;53:200-220.
- Price PI, Ostendorf M, Shattuck-Hufnagel S and Fong C. The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America* 1991;90:2956-2970.
- Prieto P and Torreira F. The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics* 2007;35:473-500.
- Prieto P. Tonal alignment patterns in Catalan nuclear falls. *Lingua* 2009;119:865-880.
- Prom-on S, Xu Y and Thipakorn B. Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America* 2009;125:405-424.
- Protopapas A and Lieberman P. Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America* 1997;101:2267-2277.
- Raidt S, Bailly G, Holm B and Mixdorff H. Automatic generation of prosody: Comparing two superpositional systems In: *Proceedings of Speech Prosody 2004*, Nara, Japan, 2004. 417-420.
- Redi L and Shattuck-Hufnagel S. Variation in the realization of glottalization in normal speakers. *Journal of Phonetics* 2001;29:407-429.
- Rialland A and Robert S. The intonational system of Wolof. *Linguistics* 2001;39:893-939.
- Rialland A. African "lax" question prosody: its realisations and its geographical distribution. *Lingua* 2009;119:928-949.
- Rietveld ACM and Gussenhoven C. On the relation between pitch excursion size and prominence. *Journal of Phonetics* 1985;13:299-308.
- Rump HH and Collier R. Focus conditions and the prominence of pitch-accented syllables. *Language and Speech* 1996;39:1-17.
- Rump HH and Hermes DJ. Prominence lent by rising and falling pitch movements: Testing two models. *Journal of the Acoustical Society of America* 1996;100:1122-1131.
- Rush, J. (1827). *The philosophy of the human voice*. 7th ed. J. B. Lippincott & Co., Philadelphia, 1879.
- Sadat-Tehrani N. The alignment of L+H* pitch accents in Persian intonation. *Journal of the International Phonetic Association* 2009;39:205-230.
- Sakurai A, Hirose K and Minematsu N. Data-driven generation of F0 contours using a superpositional model. *Speech Communication* 2003;40
- Sanderman AA and Collier R. Prosodic rules for the implementation of phrase boundaries in synthetic speech. *Journal of the Acoustical Society of America* 1996;100:3390-3397.
- Schafer AJ, Speer SR, Warren P and White D. Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research* 2000;29:169-182.
- Schafer D. The role of intonation as a cue in turn taking in conversation. *Journal of Phonetics* 1983;11:243-344.
- Scherer KR. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 2003;40:227-256.

- Scherer KR and Bänziger T. Emotional expression in prosody: a review and an agenda for future research In: Proceedings of Speech Prosody 2004, 2004. 359-366.
- Shami M and Verhelst W. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication* 2007;49:201-212.
- Shattuck-Hufnagel S and Turk AE. A Prosody Tutorial for Investigators of Auditory Sentence Processing. *Journal of Psycholinguistic Research* 1996;25:193-247.
- Shen XS. The use of prosody in disambiguation in Mandarin. *Phonetica* 1993;50:261-271.
- Shriberg EE and Lickley RJ. Intonation of clause-internal filled pauses. *Phonetica* 1993;50:172-179.
- Shue Y-L, Shattuck-Hufnagel S, Iseli M, Jun S-A, Veilleux N and Alwan A. On the acoustic correlates of high and low nuclear pitch accents in American English. *Speech Communication* 2010;52:106-122.
- Silverman K, Beckman M, Pitrelli J, Ostendorf M, Wightman C, Price P, Pierrehumbert J and Hirschberg J. ToBI: A standard for labeling English prosody In: Proceedings of The 1992 International Conference on Spoken Language Processing, Banff, 1992. 867-870.
- Sityaev D and House J. Phonetic and Phonological Correlates of Broad, Narrow and Contrastive Focus in English In: Proceedings of The 1st International Conference on Speech Prosody, Aix-en-Provence, France, 2002. 1819-1822.
- Sluijter AMC and Terken JMB. Beyond sentence prosody: Paragraph intonation in Dutch. *Phonetica* 1993;50:180-188.
- Sluijter AMC and van Heuven VJ. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 1996;100:2471-2485.
- Smith CL. Topic transitions and durational prosody in reading aloud: production and modeling. *speech communication* 2004;42:247-270.
- Snedeker J and Casserly E. Is it all relative? Effects of prosodic boundaries on the comprehension and production of attachment ambiguities. *Language and Cognitive Processes* 2010;25:1234-1264.
- Speer SR, Kjelgaard MM and Dobroth KM. The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of Psycholinguistic Research* 1996;25:249-271.
- Studdert-Kennedy M and Hadding K. Auditory and linguistic processes in the perception of intonation contours. *Language and Speech* 1973;16:293-313.
- Sugahara M. Conditions on post-focus dephrasing in Tokyo Japanese. In: Proceedings of The 1st International Conference on Speech Prosody, Aix-en-Provence, France, 2002.
- Sugahara M. Post-focus prosodic phrase boundaries in Tokyo Japanese: asymmetric behavior of an f0 cue and domain-final lengthening*. *Studia Linguistica* 2005;59:144-173.
- Sun X. The determination, analysis, and synthesis of fundamental frequency [Ph.D. dissertation]. Northwestern University, 2002; 2002.
- Swerts M. Prosodic features at discourse boundaries of different length. *Journal of the Acoustical Society of America* 1997;101:514-521.
- Swerts M and Ostendorf M. Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication* 1997;22:25-41.
- Swerts M. Prosodic features at discourse boundaries of different length. *Journal of the Acoustical Society of America* 1997;101:514-521.
- Taylor P. Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America* 2000a;107:1697-1714.
- Taylor PA. Concept-to-Speech Synthesis by Phonological Structure Matching. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 2000b;358:1403-1417.
- 't Hart J, Collier R and Cohen A. A perceptual Study of Intonation — An experimental-phonetic approach to speech melody, Cambridge: Cambridge University Press; 1990.
- Terken J. Fundamental frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America* 1991;89:1768-1776.

- Terken JMB and Hermes DJ. The perception of prosodic prominence. In: M Horne, editors. *Prosody: Theory and experiment*. Studies presented to Gösta Bruce, Dordrecht: Kluwer, 2000. 89-127.
- Thorsen N. An acoustical investigation of Danish intonation. *Journal of Phonetics* 1978;6:151-175.
- Thorsen NG. A study of the perception of sentence intonation — Evidence from Danish. *Journal of the Acoustical Society of America* 1980;67:1014-1030.
- Tokuma S and Xu Y. The effect of F0 peak-delay on the L1 / L2 perception of English lexical stress In: *Proceedings of Interspeech 2009*, Brighton, UK, 2009. 1687-1690.
- Trager, G.L.; Smith, H.L.: *An outline of English structure* (Battenberg Press, Norman 1951).
- Trainor LJ, Austin CM and Desjardins ReN. Is Infant-Directed Speech Prosody a Result of the Vocal Expression of Emotion? *Psychological Science* 2000;11:188-195.
- Traunmüller H and Eriksson A. The size of F0 excursions in speech production and perception. Working Papers. Lund, Sweden: Dept of Linguistics and Phonetics. 43, 1994.
- Tseng C-y. *Corpus Phonetic Investigations of Discourse Prosody and Higher Level Information*. *Language and linguistics* 2008:659-719.
- Tseng C-y, Su Z-y and Lee L-s. Prosodic Patterns of Information Structure in Spoken Discourse—a Preliminary Study of Mandarin Spontaneous Lecture vs. Read Speech In: *Proceedings of Speech Prosody 2010*, Chicago, 2010. 100446:1-4.
- Ueyama M and Jun S-A. Focus realization in Japanese English and Korean English intonation. *Japanese and Korean Linguistics* 1998;7:629-645.
- Uldall E. Attitudinal meanings conveyed by intonation contours. *Language and Speech* 1960;3:223-234.
- Umeda N. "F0 declination" is situation dependent. *Journal of Phonetics* 1982;10:279-290.
- van Santen J, Kain A, Klabbbers E and Mishra T. Synthesis of prosody using multi-level unit sequences. *Speech Communication* 2005;46:365-375.
- van Wijk C and Kempen G. A dual system for producing self-repairs in spontaneous speech: Evidence from experimentally elicited corrections. *Cognitive Psychology* 1987;19:403-440.
- Veronis J, Cristo PD, Courtoise F and Chaumette C. A stochastic model of intonation for text-to-speech synthesis. *Speech Communication* 1998;26:233-244.
- Wagner M. *Prosody and Recursion* [Ph.D. Dissertation]. Massachusetts Institute of Technology; 2005.
- Wagner M and Watson DG. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes* 25:905-945.
- Walker J. *The melody of speaking delineated*: (printed for the author, London 1787; reprinted by The Scolar Press, Menston. *English Linguistics* 1500–1800: No. 218, 1970). 1787.
- Wang B and Xu Y. Differential prosodic encoding of topic and focus at sentence initial position in Mandarin Chinese. *Journal of Phonetics* in press;
- Wang B, Wang L and Kadir T. Prosodic encoding of focus in six languages in China In: *Proceedings of The 17th International Congress of Phonetic Sciences*, Hong Kong, 2011.
- Wells JC. *English intonation: an introduction*, Cambridge: Cambridge University Press; 2006.
- Wennerstrom A. Intonation and evaluation in oral narratives. *Journal of Pragmatics* 2001;33:1183-1206.
- White L and Turk AE. English words on the Procrustean bed: Polysyllabic shortening reconsidered. *Journal of Phonetics* 2010;38:459-471.
- Wichmann A. Attitudinal intonation and the inferential process In: *Proceedings of The 1st International Conference on Speech Prosody*, Aix-en-Provence, France, 2002.
- Wightman CW, Shattuck-Hufnagel S, Ostendorf M and Price PJ. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 1992;91:1707-1717.
- Williams CE and Stevens KN. Emotion and speech: Some acoustical correlates. *Journal of the Acoustical Society of America* 1972;52:1238-1250.

- Wu WL and Chung L. Post-focus compression in English-Cantonese bilingual speakers In: Proceedings of The 17th International Congress of Phonetic Sciences, Hong Kong, 2011.
- Wu WL and Xu Y. Prosodic Focus in Hong Kong Cantonese without Post-focus Compression In: Proceedings of Speech Prosody 2010, Chicago, 2010.
- Xu Y. Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica* 1998;55:179-203.
- Xu Y. Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics* 1999;27:55-105.
- Xu Y. Fundamental frequency peak delay in Mandarin. *Phonetica* 2001;58:26-52.
- Xu Y. Speech melody as articulatorily implemented communicative functions. *Speech Communication* 2005;46:220-251.
- Xu Y. Principles of tone research In: Proceedings of Second International Symposium on Tonal Aspects of Languages, La Rochelle, France, 2006. 3-13.
- Xu Y. Post-focus compression: Cross-linguistic distribution and historical origin In: Proceedings of The 17th International Congress of Phonetic Sciences, Hong Kong, 2011.
- Xu Y, Chen S-w and Wang B. Prosodic focus with and without post-focus compression (PFC): A typological divide within the same language family? *The Linguistic Review* in press;
- Xu Y and Kelly A. Perception of anger and happiness from resynthesized speech with size-related manipulations In: Proceedings of Speech Prosody 2010, Chicago, 2010.
- Xu Y, Kelly A and Smillie C. Emotional expressions as communicative signals. In: S Hancil and D Hirst, editors. *Prosody and Iconicity*, forthcoming.
- Xu Y and Liu F. Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics* 2006;18:125-159.
- Xu Y and Liu F. Determining the temporal interval of segments with the help of F0 contours. *Journal of Phonetics* 2007;35:398-420.
- Xu Y and Sun X. Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America* 2002;111:1399-1413.
- Xu Y and Wang M. Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics* 2009;37:502-520.
- Xu Y and Wang QE. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* 2001;33:319-337.
- Xu Y and Xu CX. Phonetic realization of focus in English declarative intonation. *Journal of Phonetics* 2005;33:159-197.
- Xu Y, Xu CX and Sun X. On the Temporal Domain of Focus In: Proceedings of International Conference on Speech Prosody 2004, Nara, Japan, 2004. 81-84.
- Yoon T-J. Speaker consistency in the realization of prosodic prominence in the Boston University Radio Speech Corpus In: Proceedings of Speech Prosody 2010, Chicago, 2010.
- Young SJ and Fallside F. Speech synthesis from concept: A method for speech output from information systems. *Journal of the Acoustical Society of America* 1979;66:685-695.
- Zerbian S, Genzel S and Kügler F. Experimental work on prosodically-marked information structure in selected African languages (Afroasiatic and Niger-Congo) In: Proceedings of Speech Prosody 2010, Chicago, 2010. 100976:1-4.

Table 1: Hypothetical prosodic functions and their reported encoding schemes. [\[back\]](#)

Functions	Languages	Main acoustic correlates	Studies
Focus	Egyptian Arabic, English, French, Hindi, Irish, Japanese, Italian, Korean, Lebanese Arabic, Mandarin, Persian, Swedish, Tibetan, Turkish, Uygur	On-focus increase of pitch range, duration and intensity; post-focus compression and lowering of pitch range and intensity; change of pitch targets of on-focus stressed syllable (English)	Bagshaw 1993, Beyssade et al., 2009, Bolinger 1961, Botinis et al. 1999, Botinis et al. 2000, Bruce 1982a, 1982b, Chahal 2003, Chen 2006, Chen & Destruel, 2010, Chen et al., 2009, Cooper et al., 1985, de Jong 1995, 2004, de Jong & Zawaydeh 2002, Delais-Roussarie et al. 2002, Di Cristo A and Jankowski 1999, D'Imperio 2001, Dohen & Løevenbruck 2004, Dorn & Ni Chasaide, Eady et al., 1986, Gerard & Dahan 1995, Face 2006, Féry & Kügler 2008, Heldner 2003, Heldner & Strangert 2001, Hellmuth 2006, Ipek 2011, Jin 1996, Kabagema-Bilan 2011, Lecumberri 1997, Lee 1956, Lee & Xu 2010, Nootboom & Krut 1987, Patil et al. 2008, Rump & Collier 1996, Sadat-Tehrani 2009, Sugahara 2002, 2005, Ueyama & Jun 1998, Wang et al. 2011, Xu 1999, Xu & Xu 2005, Xu et al. 2004,
	Cantonese, Buli, Chichewa, Chichewa, Chitumbuka, Dagbani, Deang, Durban Zulu, Ewe, Hausa, Konni, Northern Sotho, Qiang, Taiwanese, Tangale, Vietnamese, Wa, Wolof, Yi, Yucatec Maya	<i>Lack of post-focus compression of pitch range and intensity</i>	Chen et al. 2009; Downing 2008, Gu & Lee 2007; Hartmann 2008, Hartmann & Zimmermann 2007, Jannedy 2007, Kügler & Skopeteas 2007, Nguy et al. 2008, Pan 2007, Riialand & Robert 2001, Schwarz 2009, Wang et al. 2011, Wu & Chung 2011, Wu & Xu 2010, Xu 2011, Zerbian et al. 2010
Different types of focus	Dutch, German, English, Italian	<i>No consistent differences found between narrow and corrective focus in Dutch, and between narrow and contrastive focus in English</i>	Avesani & Vayra 2003, Baumann et al. 2007, Hanssen et al. 2008, Sityaev & House 2002
Prominence	Dutch, English	Similar to on-focus cues; but no explicit examination of post-focus cues	Calhoun, 2010, Gussenhove et al. 1997, Hermes & Rump 1994, Kochanski et al. 2005, Ladd et al. 1994, Rietveld & Gussenhoven 1985, Rump & Hermes 1996, Terken 1991, Terken & Hermes 2000
Newness vs. givenness	Dutch, English, German, Mandarin	<i>Lack of consistent acoustic correlates independent of focus</i>	Féry & Kügler 2008, Most & Saltz 1979, Nootboom & Krut 1987, Wang & Xu in press
Boundary marking, Grouping, structuring	Dutch, English, French, Hebrew, Korean, Mandarin, Brazilian Portuguese	Domain/group-final lengthening, F ₀ reset	Barbosa 2007, Barbosa & Bailly 1994, Beckman & Edwards 1990, Berkovits 1994, Byrd & Saltzman 1998, 2003, Campbell, 1993, Carlson et al., 2001, Cho 2004, Cho & McQueen 2005, Edwards & Beckman 1988, Ferriera 1993, Fougerson 2001, Gussenhoven & Rietveld 1992, Krivokapic 2007, Nakatani et al., 1981, Ouden et al., 2009, Redi & Shattuck-Hufnagel

			2001, Sanderman & Collier 1996, Swerts 1997, Wagner 2005, Wightman et al. 1992, Xu & Wang 2009
Topic	Dutch, English, Mandarin	Increased sentence-initial F_0	Lehiste 1975, Sluijter & Terken 1993, Smith 2004, Tseng 2008, Umeda 1982, Wang & Xu in press
Contrastive topic	—	—	<i>No experimental investigations</i>
Turn taking	Dutch, English	Map Task dialogues	Caspers 2003, de Ruiter 2006, Oliveira & Freitas 2008, Schafer 1983
Modality, question vs. statement	Danish, English, German, Japanese, Russian, Swedish	Descriptive monologues, map task dialogues	Delattre et al. 1965, Eady & Cooper 1986, Gårding 1979, 1987, Ho 1976, 1977, Lin 2004, Liu, 2010, Liu & Xu 2005, 2007a, 2007b, Makarova 2001, McRoberts et al. 1995, Studdert-Kennedy & Hadding 1973, Thorsen 1978, 1980
	Multiple languages in the Niger-Congo, Nilo-Saharan, Afro-Asiatic, Nilo-Saharan and Afro-Asiatic superfamilies	<i>Lack of final rising;</i> Falling pitch contour, sentence-final low vowel, vowel lengthening, and a breathy utterance termination	Rialland 2009
Prosody-Syntactic interaction, Syntactic disambiguation, impact on comprehension	English, Finnish, French, Georgian, German, Italian, Japanese, Mandarin	Multiple cues	Anttila et al. 2010, Barbosa 2007, Braun & Tagliapietra 2010, Birch & Clifton 1995, Bock & Mazzella 1983, Christophe et al. 2003, Cole et al. 2010, Donati & Nespor 2003, Fery et al. 2010, Grassmann & Tomasello 2010, Ishihara 2003, Kjelgaard & Speer 1999, Lehiste 1976, Schafer et al. 2000, Shen 1993, Snedeker & Casserly 2010, Speer et al. 1996
Emotion	French, English, German, Thai, Brazilian Portuguese	Multiple cues	Auberge & Cathiard 2003, Barbosa 2009, Bänziger & Scherer 2005, Baum & Nowicki 1998, Beller 2008, Breitenstein 2001, Bulut & Narayanan 2008, Chuenwattanapranithi et al. 2008, Frick 1985, Gobl & Chasaide 2003, Ladd et al. 1985, Mauss & Robinson 2009, Ohala 1996, Pell 2001, Protopapas & Lieberman 1997, Scherer & Bänziger 2004, Shami & Verhelst 2007, Trainor et al. 2000, Wennerstrom 2001, Wildgruber et al. 2005, Williams & Stevens 1972, Xu & Kelly 2010
Attitude	Dutch, English, French, German, Japanese, Swedish	Multiple cues	Ambrazaitis 2005, Chen et al. 2004, House 2005, Morlec et al. 2001, Mozziconacci 2001, Ofuka et al. 2000, O'Shaughnessy & Allen 1983, Uldall 1960, Wichmann 2002

Table 3: Types of spontaneous speech and labeling schemes used in a number of studies. [\[back\]](#)

Study	Language	Type of spontaneous speech	Prosodic transcription/annotation
Bruce & Touati, 1992	Swedish, French	Restricted samples from conversations, interviews, political debates and radio programs	Accentual prominence; phrasing; pitch range; boundary tones; pausing
Campbell, 2004	Japanese	Phone dialogue, daily conversation	Speaker state, speaking style, voice type
Caspers, 2003	Dutch	Map task dialogues	ToDI (http://todi.let.kun.nl/), transition type
Edlund & Heldner, 2005	Swedish	Map task dialogues	Turn types
Grønnum, 2009	Danish	Descriptive monologues, map task dialogues	Pitch relation between stressed and immediate post-tonic syllable; phrasal intonation contour
House, 2005	Swedish	Human-machine dialogues	Presence of final rise and final focal accent
Kohler, 1997	German	Unspecified	Kiel intonation model (KIM) with markers of 10 prosodic domains
Nakajima & Allen, 1993	English	Map task dialogues	Discourse structure markers in terms of topic boundary classes
Ogden & Routarinne, 2005	Finnish	Phone calls and face-to-face conversations	Overlapping talk, pause location & duration
Shriberg & Lickley, 1993	American & British English	Human-machine dialogues	Filled pauses and surrounding f0 values
Swerts & Ostendorf, 1997	English	Human-machine dialogues	Discourse segmentation, utterance purpose
Swerts, 1997	Dutch	Monologues describing paintings	Boundary tones, boundary strength
Terken, 1984	Dutch	Instruction monologues	Pitch accents based on perceptual relevant F ₀ movements
Tseng et al., 2010	Mandarin	Classroom lectures	Discourse boundaries and phrasing units, emphasis
van Wijk & Kempen, 1987	Dutch	Picture description task	Self-repairs types

Table 4: Predictors (input features) used in various studies modeling pitch contours and duration. [[back](#)].

Study	Language	Predictors (input features)	Output
Bailly & Holm 2005	French	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, lexical tones in Mandarin	F ₀ , duration
Bellegarda et al. 2001	American English	ToBI transcription 40 factors, unspecified	F ₀
Barbosa 2007, 2009	Brazilian Portuguese	Position and magnitude of underlying phrase stress, and a set of dynamical control parameters	Duration
Kochanski et al. 2003	Mandarin	Lexical tone, prosodic strength, position in word	F ₀
Fletcher & McVeigh 1993	Australian English	Number of phonemes, nature of syllabic peak, position in word, position in intermediate & intonational Phrase, degree of stress, grammatical function of word, position in foot	Duration
Fujisaki et al. 2005	Mandarin	Syllable duration, amplitude of preceding tone command, which constrain timing and amplitude of tone command	F ₀
Mixdorff & Jokisch 2001	German	Boundary depth, strength, nucleus schwa/non-schwa, type of intoneme, part-of-speech, phrase index in sentence, number of phones in syllable onset, number of phones in syllable rhyme, duration of preceding phrase, <i>amplitude</i> of preceding phrase command, duration of current phrase, distance from preceding phrase command, coda voiced	F ₀
Mohler & Conkie 1998	American English	43 unspecified features, including accent type	F ₀
Morlec et al. 2001	French	Declarative, question, exclamation, incredulous question, suspicious irony, obviousness, Inter Perceptual Centre Group ratio (IPCG_ratio)	F ₀ , duration
Ni et al. 2006	Mandarin	Tone, sentence modality	F ₀
Prom-on et al. 2009	American English, Mandarin	Tone, lexical stress, focus, position in sentence	F ₀
Raidt et al. 2004	French, German	SFC: Utterance level — modality and prosodic attitude at the utterance level; syntactic structure mathematical operators IGM: 14 input parameters: Syllable level — nature of the phones included in the components of the syllable (syllable, onset, rhyme); word level — accentuation and the part of speech; phrase level or higher — neighboring boundaries and composition of current unit	F ₀
Sakurai et al. 2003	Japanese	Position of accentual phrase in utterance, number of morae, accent type, number of words, part of speech of first and last words, conjugation of first and last words	F ₀
Sanderman & Collier 1996	Dutch	Perceived boundary strength	Duration, stylized F ₀ contour

Sun 2002	American English	Vowel type, coda type, syllable stress, syllable position in word, number of syllable in word, pitch accent type, phrase accent type, number of syllables from major phrase break, part of speech, word position in sentence, number of words from major phrase break	F ₀
van Santen & Shih 2000	American English, Mandarin	No. of phonemes in the syllable, nature of syllabic peak (tense / lax vowel / diphthong / sonorant consonant), Position of syllable in foot, Position of syllable in phrase and clause, Stress assigned to syllable, and nature of pitch movement, Function/content role of the parent word	Duration

Table 5: Degrees of freedom (number of free parameters to be specified) of various models at the local level. [\[back\]](#)

Model	Study	df	Free local parameters
SFC	Bailly & Holm 2005	5	3 F_0 values per vocalic nucleus, 1 lengthening factor, Temporal scope of function
Stem-ML	Kochanski et al. 2003	8	Tone template (5 pitch values), word strength, position of template relative to syllable, length of template relative to the syllable
Fujisaki	Fujisaki et al. 2005	5	Amplitude of 1st command of syllable, amplitude of 2nd command of syllable, onset time of 1st command of syllable, end time of 1st command (and onset of the 2nd command if the second command exists) of syllable, end of 2nd command if 2nd command exists
Tone transformation	Ni & Hirose 2000, Ni et al. 2006	15	F_0 peaks for each tone (2 parameters, bottom and top frequencies of voice register of speaker, bottom and top values of voice register on the RONDO scale, damping ratio of forced vibration, peak coordination (2 parameters), parameters controlling rising characteristics of tone (2), parameters controlling falling characteristics of tone (2)
Sagging transition	Pierrehumbert 1981	4	F_0 topline (height, slope), F_0 baseline (height, slope), height of F_0 turning point (target value), time of F_0 turning point
qTA	Prom-on et al. 2009	3	Target height, target slope, rate of target approximation
Target approximation	Sun 2002	3	Target height, target slope, rate of target approximation
Tilt	Taylor 2000	5	Tilt, tilt amplitude, tile duration, tilt alignment, syllabic position