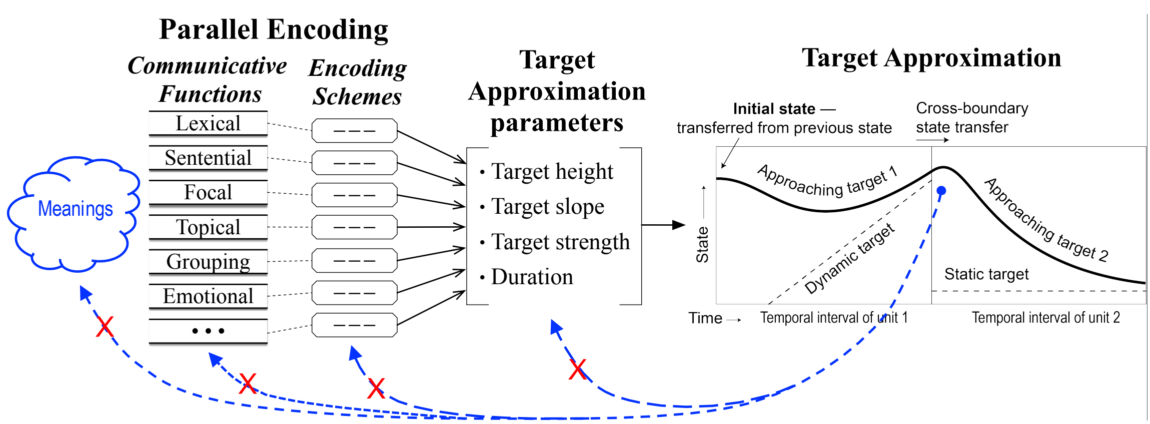
Author Response to the Commentary:

Multiple Layers of Meanings Can Be Linked to Surface Prosody without Direct Mapping

Yi Xu, Santitham Prom-on & Fang Liu

1 PENTA is not a direct mapping model

We are delighted to see Pierrehumbert's characterization of PENTA as a third generation of model of prosody and intonation. Indeed, much of the refinement PENTA may potentially bring to our understanding of prosody has benefited from knowledge gained from empirical research since the earlier models. One of the key insights from empirical findings is that surface prosodic forms, such as F_0 peaks, valleys, elbows, whole contours, etc., *cannot* be mapped to underlying units, be it tone, stress, pitch accents or prominence. This insight is instrumental in the conceptualization of PENTA, and is expressed explicitly in the presentation of the model. Figure 1 is a reproduction of the schematic of PENTA, now with the addition of potential mappings (indicated by curved arrows) to various underlying levels that are more direct than those actually assumed in the model. Also added is a representation (the cloud on the far left) of *all* the meanings that could be potentially, but not necessarily, conveyed by speech. As indicated by the crosses, not only cannot surface prosody (solid curve on the far right) be mapped directly to meanings (longest curved arrow), but also it cannot be directly linked to communicative functions, encoding schemes, underlying articulatory targets, or even the target parameters, as represented by the increasingly shorter arrows. In fact, at least three degrees of separation were recognized when PENTA was first proposed: *articulatory implementation*, *target assignment* and *parallel encoding* (Xu, 2004a, b). In other words, the very premise of PENTA is that surface "phonetic outcomes" are *not* mapped directly to meanings.



It is not enough to just point out the mismatches between meaning and phonetic outcomes, of course. Rather, PENTA is about how meanings can be ultimately mapped to surface prosody through specific connection mechanisms, so that there are no missing conceptual links. This means that each of the three degrees of separation needs to be explicitly represented in the model. Very broadly, as shown in Figure 1, meanings are first conventionalized into *communicative functions*, each having an *encoding scheme* developed through many rounds of conversational interactions. The encoding schemes of all functions work *in parallel* to jointly determine *a single sequence of targets*. These targets are then *articulatorily* implemented through non-overlapping sequential target approximation to generate continuous surface acoustic events.

This conceptualization indeed deviates from what can be referred to as the "modern linguistic theories" of prosody (Pierrehumbert, this volume: XX) in various ways. In particular, two ideas offered by PENTA, which are mentioned in the main text of this chapter, are worth recapitulating. The first is that the function-form relation, as formulated by Saussure (1916), needs a major refinement. The second is that parametric representations should replace symbolic representations as the final link to surface phonetics. The following sections will elaborate on these points.

2 Why function-first?

Saussure's (1916) notion that linguistic units are unities of signified and signifier does not make it clear what to do if there are uncertainties about both the signifier and the signified. This vagueness has not been a major problem for segmental phonemes because their function is relatively straightforward: to differentiate words. Thanks to people's strong intuition about words, the only major uncertainty is whether a particular segment does or does not distinguish certain words in a particular language. In prosody, however, both the form of the contrastive units and their functions are often ambiguous, as can be seen in the lack of consensus on both after decades of research. It is thus tempting, and in fact has been tried many times, to first develop a descriptive account of easily observable surface prosodic features such as peaks, valleys, shapes, contours and overall trends (Bolinger, 1986; Crystal, 1969; Grabe, Kochanski & Coleman, 2007; 't Hart et al., 1990), with the hope that their meaning associations can be determined by further research. Likewise, units like pitch accents, phrase accents and boundary tones were originally summarized from "observed features of F_0 contours" without explicit association with meanings, as is made clear in Pierrehumbert (1980:59). Although there have been later efforts to link them to pragmatic meanings like truth condition and common ground (Pierrehumbert & Hirschberg, 1990), these units remain primarily defined by their forms, as is evident from the fact that transcriptions of pitch tracks are used as a major means of prosody analysis (Beckman, Hirschberg & Shattuck-Hufnagel, 2006; Silverman et al., 1992).

What is overlooked in these approaches is that this is *not* how segmental phonemes are determined. While it is true that "each language has a relatively small inventory of phonological units" (Pierrehumbert, this volume: XX), whether a particular segment should be considered as a phoneme has to be determined by whether it serves to make any specific lexical contrasts rather than whether it sounds sufficiently different from other segments (Swadesh, 1934). In other words, serving a highly *specific* functional contrast is the primary determinant of the phonemic status of a segment.

What may have made the segmental phonology different from prosody is what is known as *duality of patterning* (Hockett, 1960), which is the essence of phonology as a bottleneck that "helps the language learner to acquire a large vocabulary by allowing articulatory and perceptual patterns exhibited in one word to be reused in other words" (Pierrehumbert, this volume: XX). Here the key word is the *reuse* of the same phoneme in different words, e.g., the vowel /i/ in *bin*, *pin*, and *tin*, and the consonants /b/ and /n/ in *bin*, *ban* and *bun*. Note, however, the reuse is *within the same function*, i.e., lexical contrast. An appropriate comparison in prosody would be the reuse of on-focus expansion and post-focus compression of pitch range in foci at different sentence locations (Xu, Chen & Wang, 2011). But the reuse of the same phonetic feature would not work *across functions.* It would be hard to claim, for example, that because a post-focus High tone has the same pitch level as a pre-focus Low tone, the [low] feature is shared between the focus function and the lexical function. In other words, there is unlikely a function-independent phonological /Low/ floating around in its own right, because the [low] is only relative to other tones within the same lexical contrast function.

As recognized by Hockett (1960), duality of patterning is due to heavy crowding in the lexical contrast function, as the number of words that need to be encoded massively exceeds the number of possible distinct segmental categories. Prosody, in contrast, confronts a different kind of crowding, i.e., each prosodic dimension, e.g., F_0, is shared by many functions: lexical, focal, phrasal, topical, sentential, attitudinal, emotional, social-indexical, etc. To make things worse, the identity and nature of these functions are not clear, given the lack of reference in the form of words, either spoken or written. Faced with this difficulty, PENTA-based research has followed a *function-first* principle that goes beyond the simple function-form relation envisaged by Saussure. That is, the task of prosody modeling is to find out whether a particular set of meanings have been conventionalized into a communicative function, and what the encoding scheme of this function is like, in terms of how the various prosodic dimensions are utilized to encode its *internal categories*. Following this principle, observable prosodic forms are always treated as a *secondary* property, i.e., a means of encoding the function-internal categories. This is why PENTA-based studies never use prosodic transcription as a method of prosodic analysis.

3 Hypothesis testing by controlled experiments

Identifying communicative functions and their encoding schemes is by no means a trivial task. The multiple degrees of separation as depicted in Figure 1 means that not only are surface acoustic events not directly mapped to meanings, but also no two adjacent levels are linearly related to each other to allow analysis by inversion, i.e., deriving the underlying form directly from surface property. Starting from the right end of Figure 1, target approximation, implemented as a generative model in the form of qTA (Prom-on, Xu & Thipakorn, 2009), cannot be mathematically inversed to derive the underlying targets. So our modeling work has always used analysis-by-synthesis to estimate the underlying targets (Prom-on, Xu & Thipakorn, 2009; Xu & Prom-on, 2014). And even with this approach, the quality of the target estimation is correlated with the size of the training corpus. This means that it is simply impossible to derive authentic underlying targets from single utterances.

Moving leftward to the link between underlying targets and the encoding schemes, any single target is the end result of joint contributions by multiple encoding schemes, which makes it impossible to derive all the contributing encoding schemes from an estimated target, no matter how accurate the estimation may be.

Even within an encoding scheme, a large portion consists of conventions that stipulate arbitrary context-sensitive assignment of the target parameters (referred to by Pierrehumbert as "language-specific constraints" (p. XX). For Mandarin, for example, the Low tone would assume a Rising-tone-like target if it is followed by another Low tone. This means that even if a contour is correctly recognized as related to a Rising tone, the underlying morpheme could be either one with the Low tone or with the Rising tone. For English, as found in Liu et al. (2013), whether a stressed syllable is assigned a high or low-rising target depends on its position in word, focus status and the modality (question or statement) of the sentence. This again means that it is impossible to derive individual functions even from the estimated targets.

Finally, as indicated at the far left of the figure, not all possible meanings have conventionalized functions. It is therefore impossible to know, a priori, whether a potential meaning, no matter how useful it may seem (e.g., truth condition and common ground), can be mapped to a specific encoding scheme. For example, seven different types of focus have been suggested in Gussenhoven (2007). But so far, not even the two most obviously different types, namely, information focus and contrastive focus, have been demonstrated to be consistently distinct from each other in their prosodic realizations (Hanssen, Peters & Gussenhoven, 2008; Hwang, 2012; Katz & Selkirk, 2011; Kügler & Ganzel, 2014; Sityaev & House, 2003).

In the face of so many levels of indirect and non-unique mappings, the only viable method of discovering whether a potential meaning has developed a conventionalized function, and what the encoding scheme of that function is like, is *hypothesis testing by controlled experiments*. In this paradigm, both the function and the encoding schemes are treated as hypothetical, and experiments designed to systematically manipulate the functional content are performed. In the end, it is the outcome of the experiments, which often requires multiple studies, that can inform us, with various levels of certainty, the presence of a function and the internal structure of its encoding scheme. It is with this approach, for example, that it is determined that the most salient encoding feature of prosodic focus is post-focus compression (PFC) of pitch range and intensity in many languages, and that PFC is entirely absent in many other languages (Xu, Chen & Wang, 2012).

Even with controlled experiments, however, there is an issue of whether function- or form-defined units should be the target of testing. For example, when pitch accent is targeted in some controlled studies (e.g., Grabe et al., 2000; Shue et al., 2010; Turk & White, 1999), the method of elicitation is the same as those used in studies of focus, i.e., question-answer or negation paradigms (Cooper, Eady & Mueller, 1985; Eady & Cooper, 1986; Liu et al., 2013; Patil et al., 2008; Wang & Xu, 2011; Xu & Xu, 2005). Due to the presumption of pitch accents as phonological units, these studies either examine phonetic properties of the focused words only, or treat those of post-focus components as due to phrase accent or boundary tones that are independent of the nuclear pitch accents.

From the perspective of the function-first principle, pitch accents are merely a phonetic property, as they are identified by the presence of *local* F_0 peaks, valleys or movements that sound and/or look prominent, which may or may not be due to focus. For example, a prominent F_0 peak may occur at the beginning of an utterance even in the absence of an initial focus (Wang & Xu, 2011). Or, a prominent pitch movement may occur near the end of a sentence, which would, by definition, be treated as a nuclear pitch accent. But both production and perception studies have shown that these peaks would neither be always intended nor perceived as a sentence-final focus (Cooper et al., 1985; Rump & Collier, 1996; Xu & Xu, 2005). Furthermore, focus may not be always marked by an F_0 peak more prominent than that in a neutral focus sentence, as found in Turkish (Ipek, 2011). This is not surprising, because the presence of post-focus compression (which is attributed to deaccenting and/or a L- phrase accent in the AM theory) already enables successful perception of focus (Ipek, 2011; Rump & Collier, 1996; Xu, Xu & Sun, 2004). Focus, therefore, is empirically attested as a communicative function marked by multiple phonetic cues, including on-focus increase of pitch range, intensity and duration, and post-focus reduction of pitch range and intensity (Xu, 2011), with a temporal domain that expands even across a silent phrasal pause within a sentence (Wang, Xu & Ding, 2018). In contrast, pitch accent, even when seemingly obvious, is only one of such cues, which may not even be the most critical cue, because the presence of an F_0 peak later in the utterance would effectively prevent the perception of an early focus (Rump & Collier, 1996). It would therefore be difficult for PENTA to equate focus with nuclear accent in the phrase, as suggested in Pierrehumbert's commentary.

By the same token, boundary tone, as a cue to sentence modality (question vs. statement), is also only one of the phonetic markers of the contrast, rather than being a phonological unit in its own right. For American English, at least, the marking of modality involves not only a sentence-final F_0 rise or fall, but also a drastic raising or lowering of post-focus F_0 register (treated as due to an independent phrase accent in the AM theory), and a change of height and slope of all stressed syllables throughout the sentence (Liu et al., 2013).

4 Economy of representation and degrees of freedom

The kind of controlled experiments involved in typical empirical studies, however, can go only so far as identifying the functions and the gross patterns of their encoding schemes. To be able to account for the full details of surface prosody, a further step is needed to establish a form of representation that can generate real-speech-like continuous prosodic events. This ultimate goal is attempted in PENTA through *parametric representation*. In this regard, however, PENTA is often criticized for being uneconomical in representation (Arvaniti, this volume; Arvaniti & Ladd, 2009, 2015), given its insistence on a) pitch target for every syllable even if it is unstressed or bearing the neutral tone, and b) full specification of all targets in terms of not only target height (register), but also target slope and target strength, with no allowance for any underspecifications. But we fully agree with Pierrehumbert's remark that "the human cognitive system can learn very detailed patterns and often represents them with a great deal of redundancy" (p. XX). The redundancy is not only in terms of the multiple cues for any specific communicative function as discussed above, but also in terms of detailed continuous trajectories that carry massive variability due to articulatory mechanisms, dialectical differences and idiosyncrasies of individual speakers.

The solution to the redundancy problem explored in the PENTA approach, as detailed in the main text of our chapter, is *model-based parametric representation*. Model-based means that the representation is meaningful only with respect to a specific computational model. Parametric means that targets are specified by numerical parameters rather than symbolic features. The representation of F_0, for example, is by numerical specifications of target height, target slope and target strength, as shown in Figure 1. The parameter values are obtained neither by transcription nor by direct acoustic measurement, but by training the computational model on real speech data. Depending on the nature of the training data, the learned targets can be language-, dialect- or speaker-specific. Our computational studies so far have shown that the approach is able to generate pitch contours that are both natural sounding and functionally contrastive (Prom-on et al., 2009; Xu & Prom-on, 2014). And our pilot results based on speech corpora that are less well controlled than typical experimental data have also been encouraging.

Overall, whether a representation is sufficiently economical cannot be measured by the number of representational units assumed by a theory, but by the total amount of specifications needed to generate detailed continuous prosodic events that resemble those of natural speech. If a unit is specified only in terms of H or L, as is the case with pitch accents, phrase accent and boundary tones, somewhere down the line there have to be specifications of the exact pitch height, the onset time and offset time of the unit, and how exactly the unit is connected to adjacent units. If underspecification is assumed, sooner or later there has to be a mechanism to generate surface acoustics for the underspecified units. Without including all these specifications, it is impossible to compare degrees of freedom between different models.

Another way of assessing the economy of a model is to see how many redundant parameters are required. PENTA uses only three free parameters: height, slope and strength of targets. None of them is redundant, because they are all independently motivated. Target height is motivated by its universal recognition; target slope is motivated by the consistency of final velocity in dynamic tones (Wong, 2006; Xu, 1998); and target strength is motivated by the sluggish realization of a mid target in the neutral tone in Mandarin (Chen & Xu, 2006) and unstressed syllable in English (Xu & Xu, 2005). In comparison, the equivalent of target strength in the Fujisaki and the Task Dynamic models (stiffness) is mostly fixed (Fujisaki, 1983; Saltzman & Munhall, 1998), and so is largely redundant. On the other hand, the temporal domain of target approximation is fixed to the entire syllable in PENTA (Xu & Prom-on, 2015), so that there is virtually no temporal degrees of freedom. This also contrasts with the Fujisaki model (Fujisaki, 1983) and articulatory phonology/task dynamic model (Browman & Goldstein, 1992; Saltzman & Munhall, 1989), where the onset and offset of the commands and gestural scores are free parameters, which means much more degrees of freedom in the temporal domain than PENTA. Given that the AM theory has no strict specifications of tonal alignment, it would also face the problem of degrees of freedom in the temporal domain.

5 Conclusion

PENTA is part of an effort to develop a new way of conceptualizing the mapping between meanings and continuous acoustic signals in speech, starting from the prosodic aspect. The multi-fold complexity of prosody has forced us to go back to the first principles to reconsider the phonetic-phonology interface in light of the function-form dichotomy. As a result, PENTA is one of the most indirect models of prosody, as it explicates multiple degrees of separation between meaning and continuous surface prosody. At the same time, it also insists that there be no broken links in the theoretical conceptualization of prosody and intonation, and has implemented this tenet by proposing specific connection mechanisms in its computational implementation. What has also emerged from this effort is that model-based parametric representation could be the key to understanding not only the mapping of meaning to continuous phonetic output, but also how the acquisition of speech production is achieved (Xu & Prom-on, 2014, 2015).

References

Arvaniti, A. (this volume). The autosegmental metrical model of intonational phonology. In *Prosodic Theory and Practice*. J. Barnes and S. Shattuck-Hufnagel. Cambridge: MIT Press. pp. XX-XX.

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

Arvaniti, A. and Ladd, D. R. (2015). Underspecification in intonation revisited: a reply to Xu, Lee, Prom-on and Liu. *Phonology* **32**: 537-541.

Beckman, M. E., Hirschberg, J. and Shattuck-Hufnagel, S. (2006). The Original ToBI System and the. *Prosodic typology: The phonology of intonation and phrasing*: 9-54.

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

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica* **49**: 155-180.

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

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

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

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

de Saussure, F. (1916). Nature of the Linguistics Sign. In *Cours de linguistique générale*. C. Bally and A. Sechehaye: McGraw Hill Education.

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

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

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

Grabe, E., Post, B., Nolan, F. and Farrar, K. (2000). Pitch accent realization in four varieties of British English. *Journal of Phonetics* **28**: 161-185.

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

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

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

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

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

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

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

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

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

Pierrehumbert, J. (this volume). Comparing PENTA to Autosegmental-Metrical Phonology. In *Prosodic Theory and Practice*. J. Barnes and S. Shattuck-Hufnagel. Cambridge: MIT Press. pp. XX-XX.

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

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

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

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

Shue, Y.-L., Shattuck-Hufnagel, S., Iseli, M., Jun, S.-A., Veilleux, N. and Alwan, A. (2010). On the acoustic correlates of high and low nuclear pitch accents in American English. *Speech Communication* **52**(2): 106-122.

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

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

Swadesh, M. (1934). The phonemic principle. *Language* **10**: 117-129.

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

Turk, A. E. and White, L. (1999). Structural influences on accentual lengthening. *Journal of Phonetics* **27**: 171-206.

Wang, B. and Xu, Y. (2011). Differential prosodic encoding of topic and focus in sentence-initial position in Mandarin Chinese. *Journal of Phonetics* **39**(4): 595-611.

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

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

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

Xu, Y. (2004a). Transmitting Tone and Intonation Simultaneously — The Parallel Encoding and Target Approximation (PENTA) Model. In *Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing**:** 215-220.

Xu, Y. (2004b). The PENTA model of speech melody: Transmitting multiple communicative functions in parallel. In *Proceedings of From Sound to Sense: 50+ years of discoverries in speech communication*, Cambridge, MA**:** C-91-96.

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

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

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

Xu, Y. and Prom-on, S. (2015). Degrees of freedom in prosody modeling. In *Speech Prosody in Speech Synthesis — Modeling, Realizing, Converting Prosody for High Quality and Flexible speech Synthesis*. K. Hirose and J. Tao: Springer. pp. 19-34.

Xu, Y. and Xu, C. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics* **33**(2): 159-197.

Xu, Y., Xu, C. X. and Sun, X. (2004). On the Temporal Domain of Focus. In *Proceedings of International Conference on Speech Prosody 2004*, Nara, Japan**:** 81-84.