Phonology **32** (2015) 505–535.
doi:10.1017/S0952675715000299

Squibs and replies

*Explaining the PENTA model: a reply to Arvaniti and Ladd**

Yi Xu
University College London

Albert Lee
University of Hong Kong

Santitham Prom-on
King Mongkut's University of Technology Thonburi

Fang Liu
University of Essex

This paper presents an overview of the Parallel Encoding and Target Approximation (PENTA) model of speech prosody, in response to an extensive critique by Arvaniti & Ladd (2009). PENTA is a framework for conceptually and computationally linking communicative meanings to fine-grained prosodic details, based on an articulatory-functional view of speech. Target Approximation simulates the articulatory realisation of underlying pitch targets – the prosodic primitives in the framework. Parallel Encoding provides an operational scheme that enables simultaneous encoding of multiple communicative functions. We also outline how PENTA can be computationally tested with a set of software tools. With the help of one of the tools, we offer a PENTA-based hypothetical account of the Greek intonational patterns reported by Arvaniti & Ladd, showing how it is possible to predict the prosodic shapes of an utterance based on the lexical and postlexical meanings it conveys.

* E-mail: YI.XU@UCL.AC.UK, ALBERTLEE@HKU.HK, SANTITHAM@CPE.KMUTT.AC.TH, F.LIU@ESSEX.AC.UK.

We would like to thank Amalia Arvaniti, Antonis Botinis, Bronwen Evans, Bob Ladd and four anonymous reviewers for their comments on earlier drafts of this paper. This work received support from the following sources: the National Science Foundation (NSF BCS-1355479 to the first author), the Royal Society and the Royal Academy of Engineering through the Newton International Fellowship Scheme (to the third author) and the Thai Research Fund through a Research Grant for New Researchers (TRG5680096 to the third author).

1 Introduction

The Parallel Encoding and Target Approximation (PENTA) model of speech prosody was proposed as an attempt to improve the understanding of prosody by putting emphasis on two aspects of speech prosody that had not received sufficient attention, namely, communicative functions and articulatory mechanisms (Xu 2005). The goal was to develop a framework that would explain how speech prosody works as a system of communication. More specifically, the framework to be developed should be able to describe how prosody can enable a rich repertoire of communicative functions to be simultaneously realised by an articulatory system, so that all the details of the surface prosody can be traced back to their proper sources. This was an ambitious goal, which could not be achieved in one fell swoop. Much subsequent work has therefore been done in terms of empirical testing, theoretical elaboration and computational modelling (Liu & Xu 2005, Prom-on *et al*. 2009, Wang & Xu 2011, Xu & Liu 2012, Liu *et al*. 2013, Xu & Prom-on 2014).

PENTA has received much scrutiny since its proposal, and one of the most comprehensive critiques is offered by Arvaniti & Ladd (2009). Arvaniti & Ladd contrast PENTA with the autosegmental-metrical theory of prosody (Pierrehumbert 1980, Beckman & Pierrehumbert 1986, Pierrehumbert & Beckman 1988, Gussenhoven 2004, Ladd 2008), and argue that it is inadequate to explain the prosody of Greek wh-questions examined in their study. Such a direct theoretical comparison is welcome, as it provides an opportunity to explain PENTA in a way that is more directly relevant to phonology, as will be done in this paper. We will try to achieve this by offering not only an overview of the model, but also an illustration of how it can be applied in studying the prosody of specific languages. Along the way, we will also provide responses to Arvaniti & Ladd's specific criticisms. Finally, we will offer hypothetical interpretations of the prosody of Greek wh-questions based on data presented by Arvaniti & Ladd, with the caveat that the validity of all of our interpretations awaits rigorous empirical testing in future studies.

2 An outline of PENTA

2.1 Motivation and development

One of the greatest difficulties in studying prosody is what can be referred to as 'the lack of reference problem' (Pierrehumbert 1980, 2000, Xu 2011a). That is, due to the general absence of orthographic representations of prosody other than punctuation, which itself may be due to a general difficulty in judging prosodic meanings by native speakers, there is little to fall back on when it comes to identifying prosodic units, whether in terms of their temporal location, scope, phonetic property or

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

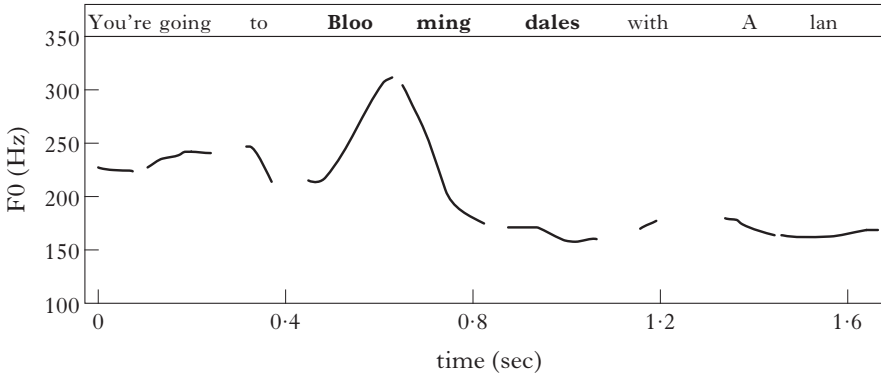


Figure 1

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

communicative function. For example, for the pitch track shown in Fig. 1, it is hard to determine what the relevant prosodic units are – F0 peaks and valleys, turning points, size of the F0 movements, temporal scope of a continuous movement, all of these, or none? The lack of reference problem makes it difficult to decide whether any of them should or should not be considered as the relevant units, and this difficulty lies at the heart of most of the theoretical disputes in speech prosody.

The strategy adopted by autosegmental-metrical theory, as best explained by Pierrehumbert (1980: 59), is to first focus on developing a formal structure of prosody by identifying which elements appear categorically distinct from each other in perception or in production. The result of such a form-first approach is the development of the autosegmental-metrical framework of prosody, which encompasses a rich inventory of phonological primitives that form the intonation systems of English (Pierrehumbert 1980, Ladd 2008), as well as many other languages (see the papers in Jun 2005). Overall, although there are variations in theoretical details and methodological approaches among studies in the autosegmental-metrical tradition, the central question in this approach remains the same (Pierrehumbert 1980, Gussenhoven 2004, Ladd 2008): what does the phonology of speech prosody look like?

The development of PENTA followed a different approach. It started with another question: how does prosody work as a communication system? Answering this question entails finding answers to two other essential questions: (i) what are the meanings that are conveyed by prosody?, and (ii) how does prosody encode these meanings in a way that allows easy decoding? In our search for answers, we took the bootstrapping strategy of always keeping one side of the function–form link relatively unambiguous while exploring the other side. In the first step, lexical tones, whose function and identity are relatively unambiguous, were experimentally examined to establish the basic mechanisms of tone

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

production in connected speech, as summarised in Xu (2005, 2011a). These studies established that even syllable-bound lexical tones do not show stable F0 properties in connected speech, but exhibit extensive surface variability according to tonal context (contrary to Arvaniti & Ladd's 2009: 65 claim that PENTA assumes stable syllable-by-syllable specification of F0 contours for tones). It was further established that articulatory inertia and tone–syllable synchrony can account for a large portion of contextual tonal variability (Xu 2005, Xu & Wang 2001). Based on findings from tone research, non-lexical prosodic functions that could be experimentally controlled were then examined, with the tonally established articulatory mechanisms as the basis for separating F0 properties that are articulatorily obligatory and those that are functionally specified (Xu 2005, 2011a). To enhance the robustness of this articulatory-functional approach, computational modelling tools were also developed as an additional, more rigorous means of hypothesis testing (Prom-on *et al*. 2009, Xu & Prom-on 2014).

Thus the PENTA approach is based on two key positions. The first is that prosodic contrasts are defined functionally, rather than by formal categories. This position touches on the fundamental issue of the role of phonology as a level of abstract representation in speech prosody. In the PENTA model, representational units are contrastive not because they are distinct from each other, but because they serve to distinguish specific functional categories (or to represent functional dimensions if they are not categorical). While this is a standard principle in phonology, the special challenge of prosody, as mentioned above, has motivated an insistence on the primacy of function in the function–form relation, especially in case of uncertainty. For example, the long-standing autosegmental-metrical debate over whether LH* and H* are distinct phonological categories in English prosody (Ladd 2008) is a non-issue in the PENTA model, since there is thus far no consensus on what functions the two tone types serve to contrast. The second position is that PENTA considers articulatory mechanisms as essential, and incorporates them into the core of its theoretical framework. In this way a large portion of the surface prosodic patterns, e.g. in terms of alignment, scaling, etc., is attributed to obligatory articulatory processes rather than to phonology.

One thing that PENTA does share with autosegmental-metrical theory is the full recognition of arbitrary rules in prosody, just as in the segmental aspect of speech. In PENTA, this recognition, which is part of the basic assumption behind the encoding schemes, is motivated (Xu 2005) by the well-known phenomenon of tone sandhi (Chen 2000). That is, the surface forms of lexical tones often vary in ways that are quite arbitrary and language-specific, and cannot be explained by clear articulatory mechanisms. PENTA assumes that similar arbitrary rules also exist in prosody; on the basis of this assumption a number of target-assignment rules which are dependent on factors like the stress pattern of words, focus and modality have been recognised for American English (Liu *et al*. 2013, Xu & Xu 2005). Some of these will be illustrated in §3.2.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

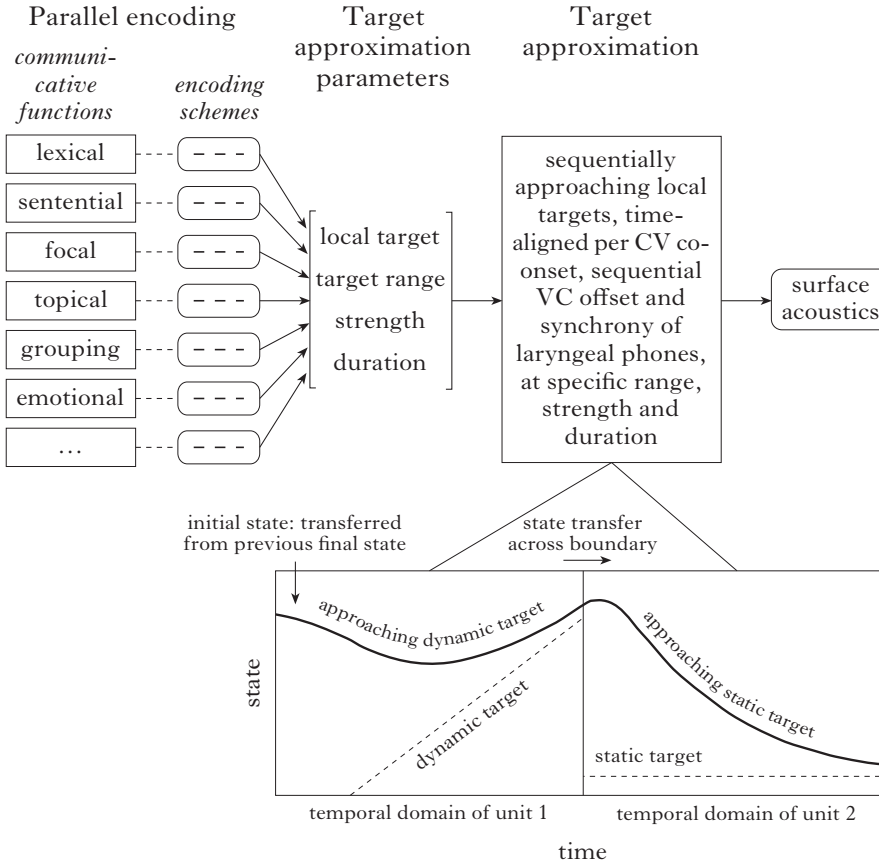


Figure 2

Upper panel: a schematic sketch of the PENTA model. Lower panel: the target-approximation component, which is an articulation process (Xu & Wang 2001, Xu 2005, Xu & Liu 2012).

2.2 The conceptual framework

Figure 2 is a schematic diagram of PENTA in its most general form, i.e. representing not only prosody, but also other aspects of speech (Xu & Liu 2012). The leftmost block in the upper panel represents communicative functions that are conveyed by speech. The functions are arranged in a stack to indicate that they are parallel to one another, i.e. with no hierarchical relations, hence the key word PARALLEL in the name of the model. The second block represents the encoding schemes associated with the communicative functions, i.e. the means to encode functional contrasts, whose schematisation here makes it clear that communicative functions do not directly control surface acoustics; rather, the two are linked through

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

specific encoding schemes. It has always been assumed in the PENTA model, though not always made fully explicit in published work, that some of the encoding schemes are highly stylised and language-specific, while others are more gradient and universal. The third block in Fig. 2 represents the target approximation (TA) parameters that are linked to the encoding schemes. These parameters in turn control the TA process represented by the fourth block. It is this articulatory process that directly generates surface acoustics, including F0, as represented by the fourth block.[1] The TA model, as depicted in the lower panel, assumes that each syllable is assigned an underlying target that has not only a height (or position), but also a slope specification. The surface F0 is the result of asymptotic approximation of the target in full synchrony with the syllable. At the boundary between two adjacent syllables, the final articulatory state of the first syllable is transferred to the second syllable. Such transfer often results in a delay of the apparent alignment of an F0 turning point, as shown in the lower panel.

There are many implications of PENTA that may not be immediately obvious from the descriptions given above, and this has often led to confusion about the model. It will thus be helpful to set out some of the most critical implications of PENTA, first in the brief list below, followed by further elaborations in the subsequent sections.

(i) Syllable-sized pitch targets are the prosodic primitives of PENTA, and as such bear the closest resemblance to tones in autosegmental-metrical theory. They differ from the autosegmental-metrical tones in that their link to surface F0 trajectories is via syllable-synchronised sequential target approximation. In contrast, linear or sagging interpolation between specified targets proposed by Pierrehumbert (1980) is the mechanism assumed in autosegmental-metrical theory, as is made clear in Arvaniti & Ladd. In the PENTA model, as shown in Fig. 2, all targets are virtual, because they do not directly correspond to observable features such as turning points, elbows or plateaus.

(ii) There are no specifications for the temporal alignment of turning points or elbows. Rather, all observed alignments are assumed to be the result of syllable-synchronised realisation of underlying pitch targets.

(iii) For each syllable, a unique target is assigned as a result of the interaction of all the communicative functions involved (as indicated in Fig. 2 by the single arrow between the target approximation parameters block and the target approximation block *vs*. the multiple arrows between the parallel encoding block and the target approximation parameters block). Thus the encoding schemes of all the functions involved jointly determine a unique target of each syllable for a particular phonetic dimension. This

[1] As postulated in Xu & Liu (2006, 2012) and recently tested in Prom-on *et al*. (2013), the notion of underlying targets applies not only to F0, but also to other properties such as vocal tract shapes for consonants and vowels, and phonation types associated with lexical, intonational or emotional functions, and their articulation follows the same dynamic principles as tone and intonation.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

INTEGRATED target therefore carries information about all the encoded functions.[2]

(iv) In contrast to its explicit assumption about articulatory mechanisms, PENTA does not explicitly stipulate a predefined inventory of communicative functions or their encoding schemes for any language. Rather, it assumes that encoding schemes, whether language-specific or universal, or categorical or gradient, have to be established experimentally by directly controlling communicative functions.

(v) Despite the assumption of a direct link between encoding schemes and communicative functions, PENTA does not directly link communicative functions to surface prosody. Rather, it assumes that communicative functions are linked to surface prosody both through articulatory mechanisms that are universal and through encoding schemes that are either universal or language-specific.

(vi) PENTA has no phonetic implementation rules that are not based on explicit articulatory mechanisms. As will be discussed in §2.3, some of the phonetic implementation rules in autosegmental-metrical theory can be reinterpreted from the PENTA perspective as being morphophonological rather than phonetic. As such, they are treated as properties of relevant encoding schemes.

To summarise, the only obligatory melodic primitives in PENTA are the syllable-sized pitch targets. The phonetic characteristics of these targets include height, slope and rate of approximation. These characteristics can be used to describe their phonetic types, such as targets that are high or low, dynamic or static (having flat or non-flat slopes), or strong or weak (having a high or low rate of approximation). As a result, although PENTA does not stipulate an inventory of predefined phonological categories, once a particular function in a language is identified, it is possible to discuss the correspondence of the PENTA-based targets with categories predefined in other theories, such as H, H*, L, L* in autosegmental-metrical theory.

2.3 Recent new conceptual developments

There has recently been a further development in the conceptualisation of the encoding schemes in PENTA (Liu *et al*. 2013). This was driven by the recognition that some of the encoding schemes of prosodic functions bear a strong resemblance to lexical morphemes, in three critical ways. First, like lexical morphemes, each of these encoding schemes consists of multiple prosodic components, which are meaningless by themselves, but act jointly to mark both intra- and inter-functional contrasts. Second, similar to lexical morphemes, an encoding scheme for a prosodic function may have allomorph-like variants, whose occurrence is conditioned by factors

[2] Note that this is different from the Fujisaki model, which assumes two separate underlying commands – accent commands and phrase commands – each generating a string of F0 contours which are mathematically combined at the final stage of the model computation to form the ultimate surface F0 contours.

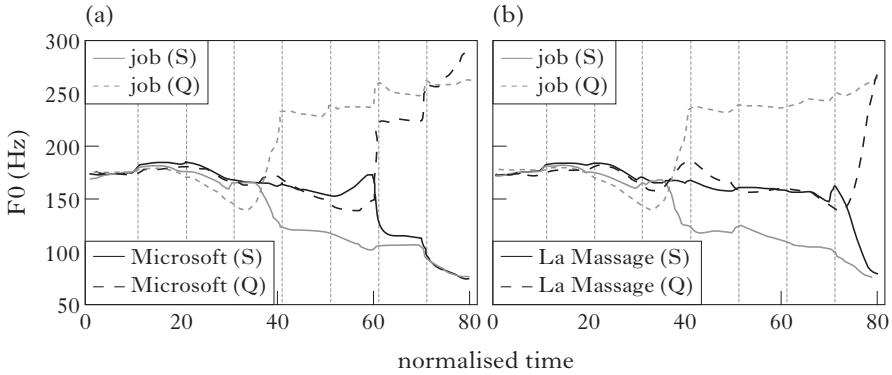


Figure 3

Mean F0 contours of focused words in statements (S) and questions (Q) of two American English sentences: (a) *You want a job with Microsoft*; (b) *You want a job with La Massage*. Data from Liu *et al*. (2013).

like location in sentence and interaction with other prosodic functions. Finally, similarly to lexical morphemes, these encoding schemes are language-specific, and their patterns may have historical origins.[3] These prosodic encoding schemes differ from lexical morphemes in that they contrast prosodic functions that carry postlexical meanings. It is therefore appropriate to refer to them collectively as prosodic morphemes.

One of the clearest examples of a prosodic morpheme is prosodic focus, whose function is to highlight one speech unit against the rest of the sentence. Empirical studies have shown that focus is realised not only with specific pitch patterns, but also with specific patterns of duration, intensity and even voice quality (Cooper *et al*. 1985, Heldner 2003, de Jong 2004). Also, in many languages, focus is realised not only with prosodic patterns of the focused unit itself, but also with POST-FOCUS COMPRESSION of pitch and intensity (see Xu *et al*. 2012 for a review). Furthermore, post-focus compression has recently been found to be absent in many other languages (Xu *et al*. 2012). It has been hypothesised that post-focus compression as a special way of encoding focus is a feature inherited from a proto-language (Xu 2011b). Thus the encoding scheme of focus in languages like Mandarin and English are multi-componential, language-specific, and probably with historical etymologies – very similar to lexical morphemes.

Another example can be found in American English, where the underlying pitch target of a stressed syllable varies depending on whether the syllable is word-final or non-final, whether the word is focused and whether the sentence is a statement or yes-no question (Liu *et al*. 2013), as can be seen in Fig. 3. Figure 3 also shows that the F0 of the post-

[3] Note that these are necessary rather than sufficient properties of morphemes. For example, having a historical lineage alone does not make an encoding scheme morpheme-like. But having all three properties makes a strong case for this analogy.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

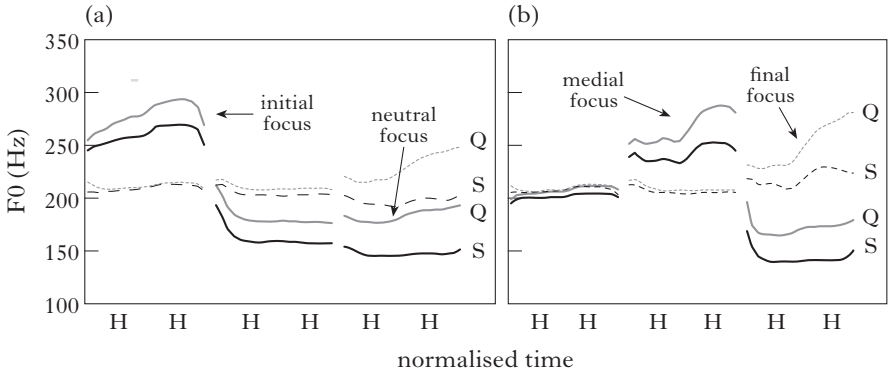


Figure 4

Mean F0 contours of the Mandarin sentence *Zhāng Wēi dānxīn Xiāo Yīng kāichē fāyūn* 'Zhang Wei is concerned that Xiao Ying may get dizzy when driving', spoken as either a statement (black lines; S) or a question (grey lines; Q). H denotes High tone. In (a), focus is on the sentence-initial word (solid lines) or there is no narrow focus (dashed lines). In (b), focus is either sentence-medial (solid lines) or sentence-final (dashed lines). Adapted from Liu & Xu (2005).

focus syllables varies markedly, depending on whether the sentence is a statement or question. In particular, post-focus F0 in a question is raised well above the reference level, i.e. the pre-focus F0. This pattern, however, is absent in Mandarin (Liu *et al*. 2013), as can be seen in Fig. 4. Such a cross-linguistic typological difference is again similar to the behaviour of lexical morphemes, although more research is needed to further explore this phenomenon.

The notion of prosodic morpheme is an alternative to the tonal morpheme proposed by Pierrehumbert & Hirschberg (1990). As discussed in detail in Liu *et al*. (2013), many of the morpheme-like meanings proposed by Pierrehumbert & Hirschberg for the phonological intonational components are similar to those associated with prosodic functions like focus and modality. But the multi-componential coding of the prosodic functions demonstrated by empirical studies show that it is these functions, rather than the pitch accents, phrase accents and boundary tones, that bear the most resemblance to lexical morphemes. Furthermore, some proposed phonetic implementation rules in autosegmental-metrical theory (Pierrehumbert 1980, Pierrehumbert & Hirschberg 1990) are part of the morpheme-like characteristics of focus and modality. For example, the upstep rule in English, which is said to raise the portion of F0 corresponding to a high boundary tone H% relative to the preceding H- phrase accent, is shown to be part of a continuous upshift of post-focus pitch range to mark a question (Fig. 3). Thus this extra raising is morphophonological, i.e. being part of a prosodic morpheme, rather than being a phonetic implementation rule.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

2.4 A quantitative realisation

Like most other theoretical intonation models (O'Connor & Arnold 1973, Pierrehumbert 1980, Bolinger 1986, 't Hart *et al*. 1990), PENTA was qualitative at the time of its proposal (Xu 2005). As such, it could be employed in qualitative description and explanation of speech data, hypothesis testing and making qualitative predictions, but could not be used to make numerical predictions about intonation. An early effort was made to quantify the TA model (Xu *et al*. 1999), followed by a much improved implementation in the form of the QUANTITATIVE TARGET APPROXIMATION (qTA) model, which also enabled full testing of PENTA (Prom-on *et al*. 2009). The development of qTA followed a number of principles. The first was that there should be as few free parameters as possible, and every free parameter should be meaningful, i.e. usable by one or more encoding schemes. The second principle was that all the critical components of TA described in §2.2 should be quantitatively implemented, so as to faithfully realise the theoretical model. The third principle was that the model parameters should be learnable from real speech data, so as to enable full-fledged numerical testing of the predictive power of the theoretical model.

In qTA, the F0 of each syllable is represented by a third-order critically damped linear system driven by a pitch target, as shown in (1), where the first term represents the pitch target as a straight line with slope m and height b. The second term represents the natural response of the system, in which the transient coefficients, c_1, c_2 and c_3, are calculated based on the initial F0 dynamic state and pitch target of the current syllable. As such they are not free parameters. Parameter λ represents the strength of the F0 movement toward the target. qTA realises the state transfer between adjacent syllables by taking the final F0 state of the preceding syllable in terms of its final F0, $f_0(0)$, velocity, $f_0'(0)$, and acceleration, $f_0''(0)$, as the initial F0 dynamic state of the current syllable.

(1) $f_0(t) = (mt + b) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t}$

With this initial state the three transient coefficients are computed with the formulas in (2).

(2) a. $c_1 = f_0(0) - b$
b. $c_2 = f_0'(0) + c_1\lambda - m$
c. $c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2$

Thus, for each syllable, qTA has only three free parameters: m, b and λ. m and b specify the form of the pitch target, with positive and negative values of m indicating rising and falling targets, and positive and negative values of b indicating raised and lowered pitch targets relative to the speaker's average F0. λ indicates how rapidly a pitch target is approached, with higher values representing faster target approximation.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

qTA therefore provides a faithful numerical representation of all the critical aspects of the theoretical TA model.

The development of the TA model and its qTA implementation were inspired by empirical findings about tonal dynamics (Xu & Wang 2001, Prom-on *et al*. 2009), and were independent of other models, although similarities to a number of existing quantitative models became clear *post facto*. Despite the similarities, however, at least three key features remain unique to qTA on close examination: (i) unitary dynamic targets (which are different from contour targets, as in the Stem-ML (Kochanski & Shih 2003) and SFC (Bailly & Holm 2005) models), (ii) unidirectional sequential target approximation, i.e. no overlap of movements, as in the task-dynamic model (Saltzman & Munhall 1989), or return phase in a movement, as in the Fujisaki model (e.g. Fujisaki 1983), (iii) high-order state transfer across target approximation movements, a feature not found in any other model except VocalTractLab, which adopts the same idea and makes the transfer order even higher (Birkholz *et al*. 2011).

2.5 Why is there a pitch target for every syllable?

One of the most questioned aspects of PENTA is its assumption of a pitch-target specification for each syllable in any language. This might appear to be an overgeneralisation from a tone language, and gives the impression of overfitting for languages that are not lexically tonal. In English and Greek, for example, many syllables appear unspecified for pitch because of their high F0 variability, absence of prominent peaks or valleys, and lack of stress. It therefore seems natural to assume, as does autosegmental-metrical theory, that 'not every syllable has to have a specification for pitch' (Arvaniti & Ladd 2009: 48). Similar 'sparse tonal specification' assumptions can be found in other models as well (e.g. Fujisaki 1983, Hirst 2005).

PENTA's imperative for pitch target for each syllable comes from its core assumption about speech articulation, as represented by the TA model shown in Fig. 2. That is, the F0 contour of every syllable comes from a single mechanism: articulatory approximation of an underlying pitch target in synchrony with the syllable. Thus there is no other way of generating an F0 contour for a syllable than assigning it an underlying pitch target. It is possible, however, to allow a single pitch target to be assigned to a string of unstressed syllables, as in the Fujisaki model. There are two reasons why we choose not to do so. The first is our assumption that the syllable, as a basic coarticulatory unit, is produced with all its underlying targets fully specified, whether consonantal, vocalic or tonal, and the process of articulation is to realise all of them simultaneously through target approximation within a time structure provided by the syllable (Xu & Liu 2006, 2012). In other words, because all the targets, including the pitch target, have to be articulated in coordination at the syllable level, it is impossible for surface F0 contours to be generated separately and then added to the syllable. The second

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

reason is that there is evidence, as will be discussed later, that not only stressed syllables but also unstressed syllables are assigned function-based contrastive pitch targets. For example, Xu & Xu (2005) found that when an initial-stressed word in English was focused, any unstressed syllables were assigned post-focus targets, i.e. with actively lowered pitch. But an unstressed syllable is also assigned a low strength, which is consistent with its weak stress status. As found in both acoustic analysis (Xu & Xu 2005, Chen & Xu 2006) and computational modelling (Liu *et al*. 2013, Xu & Prom-on 2014), such low strength can account for the high variability (and hence an apparent lack of target) of the pitch of the unstressed syllables in English and the neutral tone in Mandarin. Also, as will be shown later, similar differential strength assignments can, at least hypothetically, account for the alignment patterns in Greek wh-questions reported by Arvaniti & Ladd.

As further support, there is evidence that computational models with F0 specifications for every syllable generate synthetic prosody with better numerical and perceptual quality than those that have non-syllabic pitch specifications (Sun 2002, Raidt *et al*. 2004). Sun (2002), in particular, found that the three-target model (Black & Hunt 1996), which simply uses three F0 points for each syllable, generated better synthetic prosody than did the Tilt model (Taylor 2000), which uses a sophisticated algorithm to represent the detailed shape of F0 peaks, when both models were trained on the same corpus.

Finally, in terms of economy of representation, the assumption of one target per syllable may not be as uneconomical as it appears. This is because, although each syllable needs to be assigned a target, the target can be the same for all syllables with the same functional status in terms of lexical tone, lexical stress, focus, modality (i.e. question *vs*. statement), boundary marking, etc. Such economy of representation is helped by PENTA's assumption of full synchrony of pitch targets with the syllable, which eliminates the need for parameters that represent the temporal alignment of onset and offset of prosodic units relative to segments, as is obligatory in models that assume flexible timing (Pierrehumbert 1981, Fujisaki 1983).[4] As will be shown in greater detail in §3, only a small number of target parameters are needed to represent lexical tone, lexical stress, focus and modality in English, Mandarin and Thai. With these parameters, the intonation of all utterances in the corpora of the three languages was predictively synthesised with high accuracy in terms of root mean square errors and correlations when compared to the natural F0 contours (Prom-on *et al*. 2009, Prom-on & Xu 2012).

[4] It could be argued, as pointed out by one reviewer, that there is no *a priori* reason why the temporal domains for different tasks being produced in parallel have to coincide. But a model has to have an assumption about timing, and flexible timing (as in the Fujisaki model and target-interpolation model) and fixed timing (as in the SFC model (Bailly & Holm 2005) and the three-target model (Black & Hunt 1996)) are both obvious choices.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

3 Encoding schemes and their parametric representations

The above outline of PENTA, though more detailed than in previous publications, still leaves some ambiguities about the model, especially in terms of the nature of the encoding scheme and its relation to phonological representation. For further clarification, we would like to start by reiterating the core tenet of PENTA, mentioned at the beginning of §1, which is to develop a model that can explain exactly how speech works as a communication system. Based on this, we need to understand not only how meanings are encoded, but also how the coding is done in production and perception, how it can be learned in acquisition and how it may change over time. In other words, we need to know how this system *operates*. From an operational perspective, encoding schemes are the link between the meanings to be conveyed and the articulatory processes with which they are represented, in a way that allows effective transmission to the listener. A major task in the PENTA approach is therefore to identify the encoding schemes of various communicative functions. Empirical studies following this approach have shown that many meanings are conveyed by morpheme-like encoding schemes, as mentioned earlier. But some other meanings, e.g. emotion, attitude, etc., are conveyed by encoding schemes that are less stylistic, more universal, and likely shared with other animals (Xu, Kelly & Smillie 2013, Xu, Lee *et al*. 2013). The notion of encoding scheme therefore covers both types of meanings.

The assumption that encoding schemes need to be empirically discovered means that, in principle, the repertoire of encoding schemes in PENTA is an open set. But there are also clear constraints that significantly limit the size of the repertoire. These may come from very diverse sources, however. One major source is articulatory mechanisms, some of which are built into PENTA. For example, articulatory inertia makes it impossible for F0 movements to go beyond the maximum speed of pitch change, which would exclude pitch targets whose slope is too steep. Also, syllable-synchronised target approximation means that the timing of underlying targets relative to the syllable is largely fixed. Diachronic changes are another source of constraints. For example, the cross-linguistic distribution of post-focus compression found in recent research, as discussed in §2.3, has led to the Nostratic origin of post-focus compression hypothesis, which makes strong predictions about the existence of post-focus compression in all languages (Xu 2011b). Finally, findings about emotional expressions in speech have pointed to the bio-informational principles of vocal coding that humans presumably share with other animals (Xu, Kelly & Smillie 2013, Xu, Lee *et al*. 2013). This again offers strong predictions about emotion-related encoding schemes. Given the diversity of the sources of constraint, PENTA is a framework that groups together mechanisms that are independent of one another, but treats all of them as indispensable parts of the speech-communication process.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

More importantly, the recognition of the articulatory mechanisms has also shed new light on the issue of mental representation of prosody. Given the basic tenet of the PENTA approach as mentioned above, it is imperative that the assumed mental representation is operational. This means, first, that the representation should be sufficiently abstract so as not to require too much memory space. Second, it also needs to be able to account for fully continuous surface forms, leaving as few details unexplained as possible. Third, it should allow full gradience, so as to adequately represent individual and dialectal variation. Finally, it needs to be learnable with testable computational algorithms. The solution found in the PENTA approach, as a result mainly of the efforts to develop a computational realisation of the theoretical model, is a PARAMETRIC REPRESENTATION in the form of underlying target, as opposed to symbolic representations that directly correspond to phonological units. Here the parametric representation is interpretable only on the basis of specific articulatory mechanisms that can be simulated with a computational model. For PENTA, qTA, introduced in §2.4, is such a model. Using data from English and Mandarin as examples, the next two sections will briefly show how parametric representations operate in PENTA.

3.1 Computational modelling tools

Since qTA was first proposed, computational tools have been developed to enable its conceptual exploration and quantitative testing. So far, four tools have been developed. qTA_demo1, mentioned by Arvaniti & Ladd (2009: 65), and qTA_demo2 are web-based interactive Java programs that demonstrate how the qTA model works.[5] Their interactive features make them convenient tools for a quick impromptu test of an idea or a prediction based on the TA model (as can be seen in Fig. 8, to be discussed later).

The other two tools, PENTAtrainer1 (Xu & Prom-on 2010–14) and PENTAtrainer2 (Xu & Prom-on 2014), are data-driven modelling programs.[6] Both use machine learning algorithms to automatically extract target parameters from real speech data through analysis-by-synthesis. These learning algorithms test each candidate target by putting it into the qTA function to generate continuous F0 contours that are then compared to the natural contours. The goodness of fit between the synthetic and original contours is used as the criterion in the selection of the targets (Prom-On *et al*. 2009, Prom-on & Xu 2012). The quality of the F0 generation is assessed by three means: (i) root mean square errors, which measures the discrepancy of the synthetic contours from the original contours in terms of point-by-point height difference, (ii) Pearson's *r*, which assesses how closely the overall shape of the synthetic contours

[5] Available (October 2015) at http://www.phon.ucl.ac.uk/home/yi/qTA and http://www.phon.ucl.ac.uk/home/yi/qTA_demo2 respectively.

[6] Available (October 2015) at http://www.phon.ucl.ac.uk/home/yi/PENTAtrainers.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

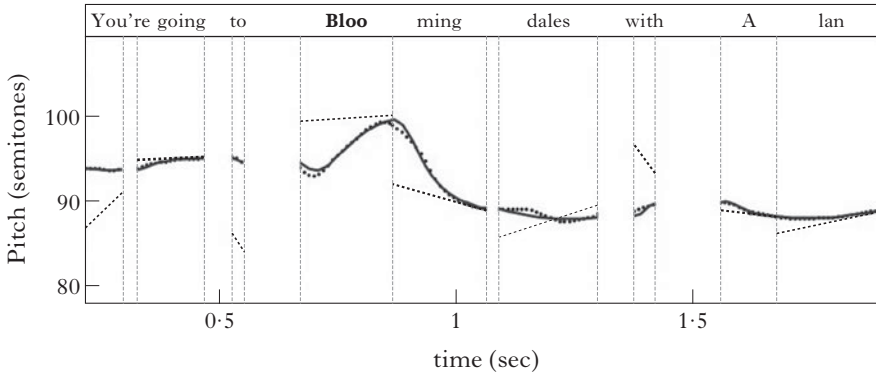


Figure 5
Original (dotted) *vs.* resynthesised (solid) F0 contours of the English utterance *You're going to Bloomingdales with Alan* shown in Fig. 1. Adapted from a synthesis by PENTAtrainer1 (http://www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer1/).

correlates with that of the original contours, and (iii) perceptual evaluation in terms of category identification (e.g. tone, focus, etc.) and naturalness.

Critically, both trainers allow predictive synthesis of F0 contours using categorical parameters learned from training. They differ only in terms of how function-specific targets are obtained. PENTAtrainer1 takes a two-phase approach. In Phase 1, an optimal target is obtained for each syllable of each utterance by comparing the performance of all possible combinations of the three target parameters (b, m, λ in (1)). The parameter set that achieves the best fit to the F0 contour of a specific syllable (i.e. with the smallest sum square errors) is selected as its pitch target. An example of such resynthesis is shown in Fig. 5, where the short dashed lines are the learned targets. The F0 contours generated with these learned targets (solid lines) seem to fit the original F0 contours (dotted lines) quite well. In Phase 2, categorical targets are obtained by averaging over the parameters of all the syllables in the corpus that belong to the same categorical combination, e.g. all the on-focus H tones that occur at the beginning of a sentence (Prom-on *et al*. 2009). This approach can be referred to as CATEGORISATION BY AVERAGING. As found in Prom-on *et al*. (2009) and Liu *et al*. (2013), good predictive results can be obtained for both English and Mandarin.

The categorisation by averaging strategy employed in PENTAtrainer1, despite its reasonable performance, cannot satisfactorily estimate all qTA parameters. In particular, locally estimated parameters may not be globally optimal. For example, in some cases, the rate of target approximation (λ) may not be adequately estimated if there is severe target undershoot. Besides, the simple exhaustive search implemented in PENTAtrainer1 is inefficient, and probably biologically unrealistic as a learning algorithm.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

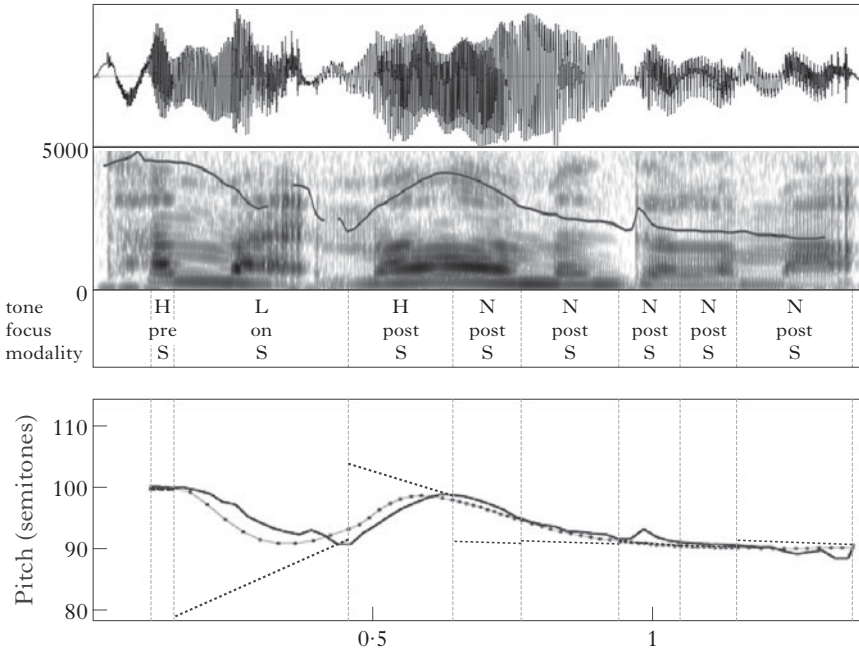


Figure 6

A schematic representation of PENTAtrainer2 interfaces (http://www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer2/) for the Mandarin sentence *tā MǍI māma men de la ma* 'Did he BUY what mother has?', with focus on *mai3*. The annotation interface (top) allows users to mark temporal scope of functional units. Here and below, the annotated functions are lexical tone (H = High, L = Low, N = neutral), focus (pre = pre-focus, on = on-focus and post = post-focus) and sentence modality (S = statement, Q = question). Vertical lines are set to coincide with syllable boundaries. The temporal scope of a functional region covers syllables with identical labels. The output interface (bottom) displays learned pitch targets (dashed lines), as well as synthetic (dotted lines) and natural (solid lines) F0 contours, and allows users to play the utterance with either synthetic or natural prosody (Prom-on & Xu 2012).

These problems are addressed by PENTAtrainer2, in which function-specific targets are learned directly from an entire corpus that has been functionally annotated (Prom-on & Xu 2012, Xu & Prom-on 2014). This is achieved with SIMULATED ANNEALING, an optimisation algorithm that performs stochastic parameter sampling to avoid local minima in parameter estimation. Figure 6 shows an example of an annotated utterance (top) and natural F0 and synthetic contours (bottom), where the latter is generated with categorical target parameters learned from an entire corpus.

In Xu & Prom-on (2014), good overall numerical results were achieved with PENTAtrainer2 for English (the same dataset tested with

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

PENTAtrainer1 in Liu *et al*. 2013), Mandarin and Thai. In Prom-on *et al*. (2009), which applied categorisation by averaging, the perceptual identification rates for tone in Mandarin and focus in both Mandarin and English were found to be similar for synthetic and natural speech. Just as importantly, synthetic prosody (in terms of F0 and duration) was heard to be just as natural as natural prosody for English, and only slightly worse for Mandarin.

Interestingly, the total number of function-specific parameters learned from the speech corpora and used in the predictive synthesis was very small. In Xu & Prom-on (2014), 78 parameters (i.e. 26 for each of the *b*, *m* and λ values) were used for 960 English sentences (consisting of 8640 syllables), 84 parameters for 1280 Mandarin sentences (consisting of 10240 syllables) and 30 parameters for 2500 Thai disyllabic phrases. The number of function-specific parameters roughly equals the number of parameters per target × the number of simulated functions × the number of function-internal categories – non-existing category combinations. This suggests that a high level of abstraction can be achieved with PENTA-based computational approaches. The abstraction level is comparable to other models, e.g. five parameters per Standard Chinese tone in the Fujisaki model (Fujisaki 1983) and four parameters per intonational event in the Tilt model (Taylor 2000).

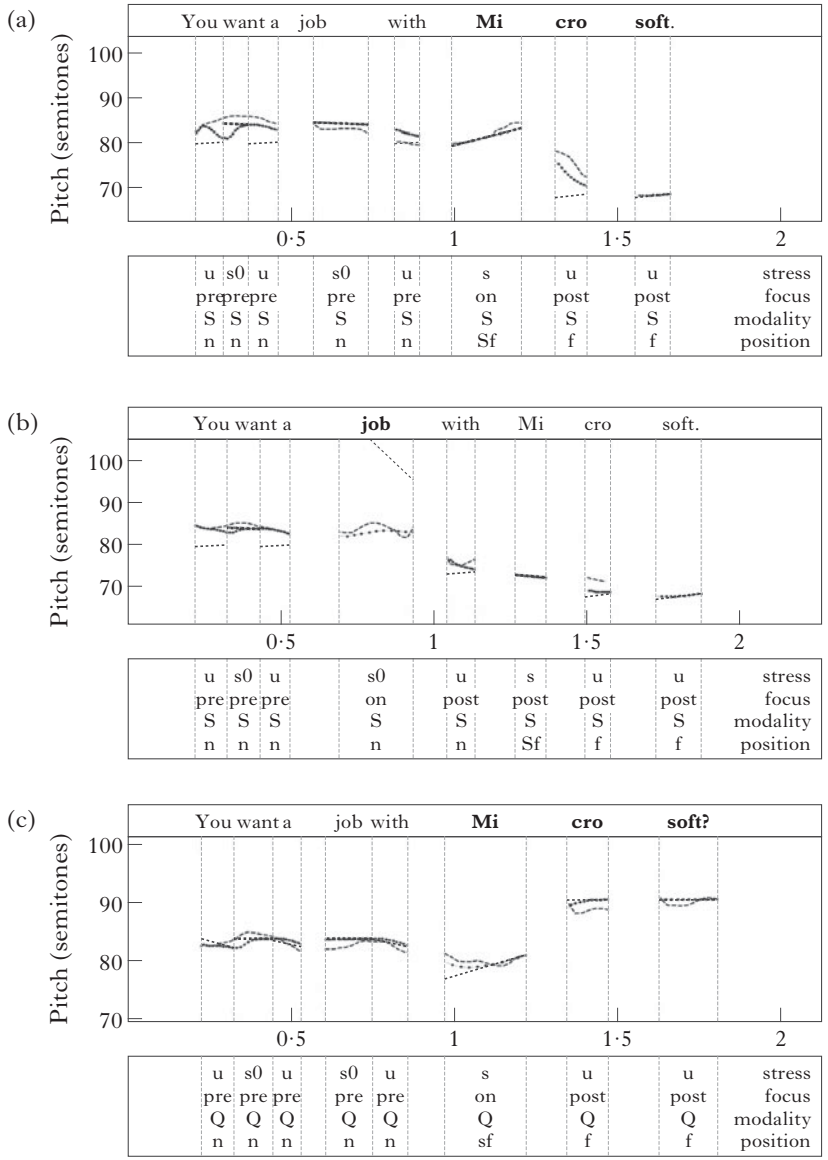
3.2 Modelling encoding schemes of English prosody: an illustration

The application of the computational tools described above allows us to model some of the major prosodic encoding schemes in English and Mandarin. Figure 7 provides a summary illustration with modelling data on English from Xu & Prom-on (2014). Each graph shows the original F0 of an American English utterance, pitch targets learned by PENTAtrainer2 and synthetic F0 contours generated with the learned targets. The sentences were spoken with either sentence-medial or sentence-final focus, either as statements or as questions. As can be seen, the encoding schemes of focus and modality in American English exhibit allomorphic patterns that are best described in terms of their interactions both with each other and with lexical stress.

(i) Focus is characterised by a robust post-focus pitch range shift, with the direction of the shift dependent on modality: downward in a statement (Fig. 7a, b), but upward in a question (Fig. 7c, d). The resulting post-focus plateaus correspond to the L- and H- phrase accents in the autosegmental-metrical model, but from the PENTA perspective they are allomorphic components of the focus and modality encoding schemes (or prosodic morphemes), rather than autonomous prosodic units in their own right.

(ii) Both focus and modality also interact with lexical stress and stress structure of the word, by determining the micro-properties of the targets. For on-focus word-final stressed syllables, the target slope falls in a statement, but rises in a question (*job* in Fig. 7b, d). For on-focus,

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press



non-final stressed syllables, the target slope rises in both statements and questions, at least for this speaker (*Mi-* in Fig. 7a, c).

(iii) In both statements and questions, targets are higher in stressed syllables than in unstressed syllables, but the differences are much smaller in questions.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

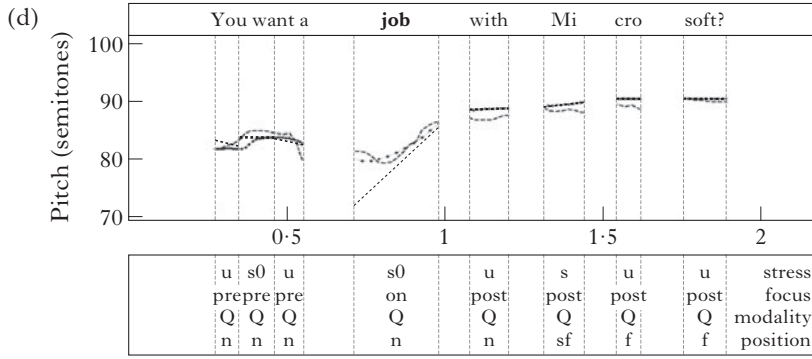


Figure 7

Original (dashed) and synthetic (dotted) F0 contours of the sentence *You want a job with Microsoft*, spoken by a male American English speaker as either a statement (a, b) or a question (c, d), with focus on either *job* (b, d) or *Microsoft* (a, c). Also displayed are the pitch targets (straight dashed lines) learned by PENTAtrainer2, based on the functional annotations shown at the bottom of each graph (stress: u = unstressed, s = non-final stressed, s0 = final stressed; syllable position (n = non-final, sf = semifinal, f = sentence-final). All graphs are adapted from screenshots of the demo window of the synthesis tool in the PENTAtrainer2 package (http://www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer2/). Data from Xu & Prom-on (2014).

In comparison with the F0 contours in Fig. 4, we can see that the variations due to cross-functional interactions in English are rather different from those in Mandarin. While English shows a robust post-focus upshift in questions, Mandarin shows a post-focus downshift even in questions, except that the size of the downshift is smaller than in statements. Again unlike English, the direction of the target slopes of Mandarin tones does not change from statements to questions, presumably due to the existence of a lexical tonal constraint. These cross-linguistic differences in the encoding schemes of similar prosodic functions show that they are highly language-specific, and that their exact forms cannot be predicted solely on functional grounds.

Note also that the match between the synthetic and original F0 contours in Fig. 7 is not nearly as good as that in Fig. 5. This is partly because the synthesis here is predictive, based on categorical parameters learned from all the utterances by a speaker in a corpus, as opposed to resynthesis in Fig. 5 (by PENTAtrainer1), but partly also because there is still room for further adjustments in the functional annotations. For example, since the relative position of unstressed syllables within an initial-stressed word is not annotated in this simulation, the pitch targets of the unstressed syllables are the same, regardless of their positions in the word. As a result, the synthetic F0 in *-crosoft* does not show final upstep in Fig. 7d. Thus, even if the major characteristics of the encoding schemes have been

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

identified, their detailed properties are still an object of continuous empirical investigations.

3.3 Model-based parametric representations

The modelling tools and the illustration of their application in the previous sections have demonstrated the plausibility of qTA-based parametric representations. These targets are functionally defined, since each of them corresponds to a unique combination of a set of functions, as shown in Fig. 6. These targets are abstract, as each of them is specified by only three parameters, but can correspond to a countless number of contextual variants. This one-to-many correspondence (Xu & Prom-on 2014) is achieved on the basis of a specific mechanistic model, namely qTA. These targets are also gradient, since all three parameters are numeric rather than symbolic. The target values are data-driven, since they are learned from real speech data. Table I displays these properties, and shows which of them are shared by symbolic representations. As can be seen, only abstractness is unquestionably shared by the two types of representations. Although it is possible to obtain autosegmental-metrical-style representations in a data-driven manner (Lee *et al*. 2014), the predictive power of doing so is as yet unknown.

	parametric	symbolic
Functionally defined	✓	
Abstract (free from redundant and variant surface detail)	✓	✓
Model-based (with mathematically defined articulatory mechanisms)	✓	
Gradient (allowing for individual and dialectal variation)	✓	
Data-driven (trainable, learnable)	✓	?

Table I

Comparison of PENTA-based parametric and autosegmental-metrical-style symbolic representations.

Model-based parametric representations may also offer a solution to a well-known puzzle in phonology, namely tone sandhi (Chen 2000). For example, Mandarin Tone 3 is changed to Tone 2 when followed by another Tone 3: T3 → T2 / __ T3. With PENTAtrainer2 this rule can be operationalised as the result of an interaction between two functions: lexical tonal contrast and boundary marking. That is, the pitch target to be implemented in articulation is jointly determined by the morphemic tone of the current syllable, the morphemic tone of the next syllable and the strength of the boundary between the two syllables. Such functional

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

interaction may allow T3 to develop a pitch-target variant that happens to be similar to that of another tone, e.g. T2. But the two need not be identical, since the functional combinations are not the same. Xu & Prom-on (2014) found that the best modelling result was obtained when the sandhi T3 was allowed to learn its own target, rather than when it was forced to use the T2 target. This result is consistent with the empirical finding of subtle yet consistent differences between the original and sandhi-derived T2 in Mandarin (Xu 1997, Peng 2000). Thus the obligatoriness of associating a unique target to each functional combination may have led to the development of tone sandhi in the first place. But further research along these lines is needed.

Finally, computational modelling of parametric representations may allow the exploration of mechanisms of speech acquisition. For example, it is known that both young songbirds and human children need to hear themselves during a critical practice stage of song or speech learning (Doupe & Kuhl 1999), but why this is the case is still unclear (Nick 2014). The analysis-by-synthesis applied in the PENTA trainers uses qTA to repeatedly generate continuous surface trajectories, and compares them to the training speech data. The ease with which near-optimal targets (i.e. those capable of predictively generating naturalistic contextual and cross-speaker variants; Prom-on *et al*. 2009, Xu & Prom-on 2014) are learned in this way suggests the importance of using one's own articulators to generate the acoustic signal during the practice period.

4 Hypothetical interpretations of Greek wh-question prosody

Because the present paper is prompted by Arvaniti & Ladd's criticism of PENTA based on their Greek data, we offer a PENTA-based interpretation of what Arvaniti & Ladd report about Greek wh-question prosody. We are not in a position to offer a full PENTA account of the Greek wh-question prosody, due to lack of experimental data on Greek, so the interpretations presented below can only be speculative, and are subject to future empirical verification.

4.1 Overall interpretation

Our overall interpretation of Greek wh-question intonation is illustrated in Fig. 8, which displays functional annotations corresponding to hypothetical underlying pitch targets and qTA-simulated F0 contours of two sentences from Arvaniti & Ladd, based on data presented in their paper. Overall, Greek wh-questions appear to involve a prosodic focus on the wh-word, which raises its pitch target(s) (the first syllable in Fig. 8a and the first three syllables in Fig. 8b), but lowers the pitch targets of all subsequent syllables. The raised on-focus pitch targets result in an early F0 peak, but the slope of the on-ramp of the peak depends on the lexical

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

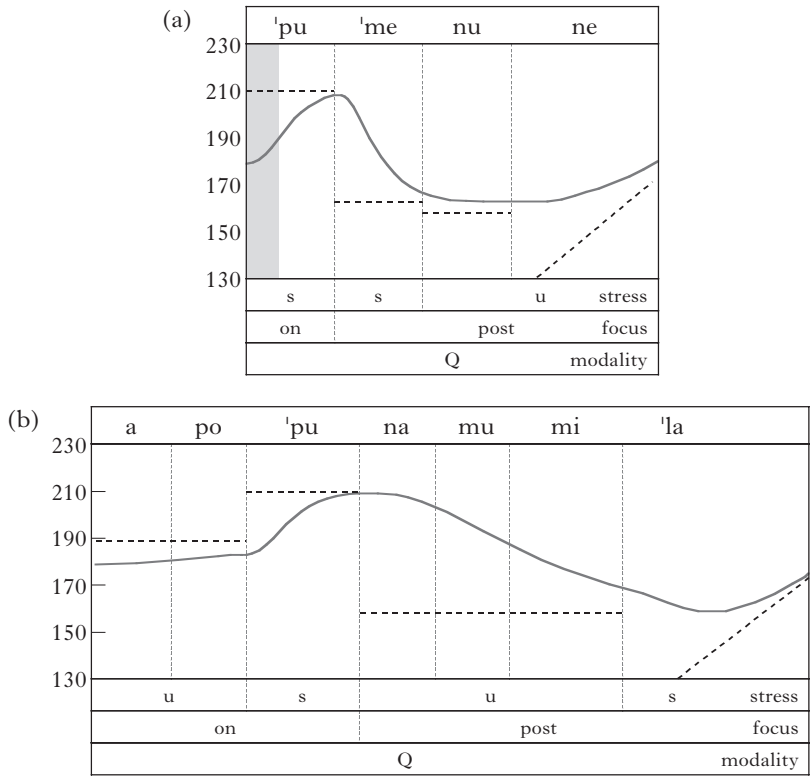


Figure 8

F0 contours, simulated using qTA_demo1, of (a) Greek ['pu 'menune] 'Where are they staying?' and (b) [apo'pu na mu mi'la] 'Where could s/he be talking to me from?', resulting from qTA realisation of underlying pitch targets (dashed lines), which are hypothetically set for the functions annotated in the bottom tiers. The vertical bar at the left edge of (a) illustrates the 'truncation' effect of a voiceless consonant.

stress of the syllable: steeper if it is stressed (a), but shallower if it is unstressed (b). The lowered post-focus pitch targets result in an F0 drop immediately after the wh-word, but the rate of the drop also depends on the lexical stress of the post-focus syllable: faster if it is stressed (a), slower if it is unstressed (b). The post-focus lowering also results in a low plateau after a post-focus stressed syllable (a). Within either the on-focus or post-focus region, the pitch target is slightly higher for a stressed than for an unstressed syllable. This is, however, purely hypothetical for Greek, and based on findings for English (Xu & Xu 2005, Prom-on *et al*. 2009, Liu *et al*. 2013, Xu & Prom-on 2014), because there is not sufficient information about stress-related target height available in the data reported by Arvaniti & Ladd. The sentence-final rise, which involves

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

a shallow rising target if the final syllable is unstressed (a), or a steep rising target if the final syllable is stressed (b), is associated with the interrogative modality of the wh-question. Overall, from the PENTA perspective, the functional equivalence of Greek wh-questions exists at multiple levels: focus shows a consistent pattern of raised on-focus pitch and lowered post-focus pitch; question modality shows a consistent sentence-final rise (or even a progressive rise throughout the sentence, if Greek is similar to Mandarin: Liu & Xu 2005); lexical stress shows (hypothetically) consistent higher *vs*. lower pitch targets. Each of these functional equivalences is shared by all the wh-question sentences presented by Arvaniti & Ladd, regardless of their length or lexical composition.

The pitch targets, represented by the dashed lines, which are purely hypothetical in Fig. 8, can be obtained by applying PENTAtrainer1 or PENTAtrainer2 to the real data. As noted above, the annotations below the F0 contours illustrate the PENTA-style functional annotations specified in the caption.

In addition to the global patterns, Fig. 8 also shows micro-patterns related to alignment, scaling, etc., which are a major concern in Arvaniti & Ladd (2009). Here we can see that they are mostly due to interactions between focus, modality and lexical stress. The details of these interactions, as will be discussed in the following sections, can be accounted for by articulatory mechanisms of pitch production, as captured by the qTA model in PENTA.

4.2 Tonal crowding, alignment and scaling: a PENTA perspective

In Arvaniti & Ladd (2009), local variations are described in terms of alignment and scaling of F0 peaks and elbows. These patterns are accounted for by tonal crowding, which is said to occur whenever two or more tones are associated with the same tone-bearing unit or with adjacent units. The evidential basis of tonal crowding is that certain observed F0 patterns vary when the phonologically specified tones are close to each other, but remain stable once those tones are two or more syllables apart. From the PENTA perspective, these tonal adjustments can be accounted for by the articulatory-functional mechanism outlined in §2.2, which involves no freedom of underlying tonal alignment, and no direct scaling as an F0 adjustment mechanism in its own right. As is shown in Fig. 8 and below, variations in both alignment and scaling can nevertheless be generated by the qTA model once the underlying pitch targets are given based on specific communicative functions.

4.2.1 *Alignment of NH as on-focus F0 peak*. NH (nuclear H) measures the location of the early F0 peak in a wh-question. Arvaniti & Ladd show that its location is earlier when the wh-word has final stress and the following word has initial stress than when there are intervening unstressed syllables, but there is no further variation according to the

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

number of intervening unstressed syllables. Also, when the interstress interval was zero, 'the peak appeared much earlier in short than in long questions, and in fact aligned with the nuclear vowel itself; in contrast, in long questions, in which the pressure on NH comes only from the following L1 [see §4.2.2 for a definition of L1], the peak co-occurred with the onset consonant of the postnuclear syllable' (2009: 58). Arvaniti & Ladd attribute these patterns to the crowding of the NH and the upcoming L, which is severe only when the L is immediately adjacent to NH.

Our interpretation, based on the TA model in PENTA and empirical data from English and Mandarin (Liu *et al*. 2013) can be seen in Fig. 8. In (a), the first post-focus syllable [me] is lexically stressed, and so its target strength is high. As a result, the rising momentum generated by approaching the on-focus high target is quickly reversed, leading to an F0 peak very close to the syllable boundary. In contrast, in (b), the first post-focus syllable is unstressed, and thus has weak target strength. As a result, it takes longer for the on-focus rising momentum to be reversed, leading to an F0 peak that is aligned further to the right of the syllable boundary. As mentioned in §2.4, evidence of such stress-related articulatory strength is found in both acoustic analysis and computational modelling for English and Mandarin. In addition, because there is no anticipatory mechanism in qTA, lexical stress of syllables further to the right would not have any more impact on the peak alignment. Thus the NH alignment reported by Arvaniti & Ladd can be accounted for by PENTA using qTA simulation without any explicit specification of F0 peak alignment or assumption of tonal crowding.

4.2.2 *Alignment of L1 as post-focus F0 elbow*. L1 refers to an elbow 'defined as the point that showed a clear change in slope between the fall after the nuclear peak and the low plateau' (Arvaniti & Ladd 2009: 55). Overall, L1 is described as exhibiting stress-seeking behaviour: it 'typically co-occurs with the first stressed syllable after the nucleus, thereby ensuring that this syllable has low F0 to the extent that tonal crowding permits' (2009: 67). From a PENTA point of view, this is directly related to the NH alignment discussed above, and thus explainable by the same mechanism. That is, as seen in Fig. 8, due to focus, F0 is lowered immediately after the stressed syllable of the wh-word, regardless of whether the first post-focal syllable is stressed. On the other hand, as also seen in Fig. 8, the speed at which this lowering is realised depends on the stress level of the post-focus syllable. It is faster if the post-focus syllable is stressed (a), but slower if it is unstressed (b). Similar stress-dependent post-focus F0 falling speed has been found for English (Xu & Xu 2005). In other words, the 'stress-seeking' behaviour observed in Arvaniti & Ladd, as well as other autosegmental-metrical-based studies (Pierrehumbert & Beckman 1988, Grice *et al*. 2000, Gussenhoven 2000), can be accounted for in PENTA as being due to the greater articulatory strength given to stressed syllables than to unstressed syllables, even when they are both post-focus.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

4.2.3 *Alignment of L2 as F0 elbow of final rise*. L2 refers to the later elbow with respect to the final vowel in a wh-question, 'defined as the point that showed a clear upward inflection between the low plateau and the utterance-final rise' (Arvaniti & Ladd 2009: 55). They found that 'in both short and long questions, L2 occurred after the onset of the final vowel, when this vowel was stressed, but slightly before it, when stress was on the antepenult; in the latter case, L2 co-occurred with the consonant of the question's last syllable' (2009: 61). More specifically, 'while L2 co-occurred with the onset of the final vowel when the last word was stressed either on the penult or the antepenult, it occurred half-way through the final vowel when this vowel was stressed' (2009: 61–62).

These patterns are again likely related to target strength due to lexical stress. That is, the target strength of sentence-final syllables is dependent on lexical stress, being higher in stressed syllables and lower in unstressed syllables. The impact of this difference can again be seen in Fig. 8. Both sentences have a sentence-final rising target associated with the question modality. The sentence in (a) shows a continuous shallow final rise, due to the low strength in its unstressed final syllable. The sentence in (b), in contrast, shows a dip in the middle of the syllable before the final rise, due to the high strength of its stressed final syllable. This dip, which is also seen in Fig. 1a in Arvaniti & Ladd (2009) for the Greek sentence ['pu 'zi] 'Where does s/he live?' with sentence-final stress, is likely to have led to the difference in the manually marked L2 alignment in Arvaniti & Ladd. But the simulation in Fig. 8 shows that the real source of the difference is in the property of the pitch targets, not in their underlying alignment.

4.2.4 *Scaling, truncation and virtual targets*. The above discussion has shown that the alignment of NH, L1 and L2 reported by Arvaniti & Ladd can be accounted for by PENTA in terms of the interaction of lexical stress with focus and question intonation. With regard to scaling, Arvaniti & Ladd did not find significant effects of tonal crowding. We note, however, that such lack of variability has much to do with the way scaling is defined, which in Arvaniti & Ladd is in terms of only the F0 peak on the wh-word and elbow of post-focus F0 drop and sentence-final F0 rise. From the perspective of target approximation, this lack of variability is not really surprising. As can be seen in the simulations in Fig. 8, this is because the time pressure is not high enough to trigger a significant undershoot for those particular measurements. For NH, there is no real leftward push from the first post-focus syllable, whether the latter is stressed or unstressed. For L1 and L2, the lack of systematic variability could also be due to a large variance in the measurement, given that visual identification of elbows is unlikely to be highly consistent. If, on the other hand, scaling refers to the degree of target undershoot in each syllable, its effect can be clearly seen in most of the unstressed syllables in Fig. 8.

Arvaniti & Ladd also report that sentences that start with a stressed syllable have higher initial F0 than those starting with an unstressed syllable. They attribute this to a truncation mechanism, by which a stressed syllable

truncates a virtual L target that occurs at the left edge of every sentence in Greek. From the simulated F0 contours in Fig. 8, however, it is difficult to see how this truncation mechanism can work. If the proposed virtual target is located at the left edge of a sentence, the stressed syllable must be at its right. Given such a target sequence, if there is any remnant of the L after the truncation, it should be still at the leftmost edge, based on the target-interpolation mechanism of the autosegmental-metrical theory, thus keeping the lowest initial F0 unchanged. With the target-interpolation model, variation of initial F0 due to stress of the sentence-initial syllable can occur only if the virtual L is fully replaced by the tone of the stressed syllable.

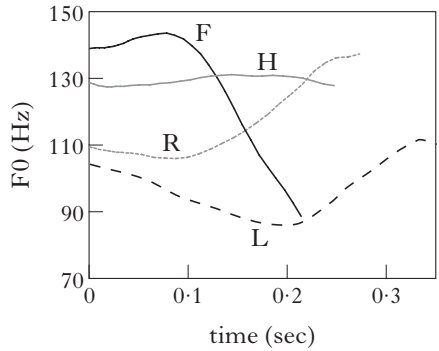


Figure 9

Mean F0 contours of Mandarin falling, rising high and low tones in the syllable [ma] spoken in isolation by 8 speakers (averaged over seven repetitions by eight speakers). Data from Xu (1997).

From the PENTA perspective, the idea of an utterance-initial virtual pitch is actually rather plausible, because there is already evidence for it in empirical data on tones produced in isolation (Xu 1997), as shown in Fig. 9. We can see that different tones have different onset F0. However, the early portions of all the tones seem to point back to a common origin in the middle of the pitch range. It is therefore possible that speakers start their laryngeal target approximation before the onset of phonation. Such a delayed voice onset is easily implementable in PENTA, by imposing a fixed time delay relative to the onset of pitch target approximation. But note that such an onset delay would 'truncate' the initial F0 from the left, rather than from the right as suggested by Arvaniti & Ladd, and would be applied regardless of whether the initial syllable is stressed.

Furthermore, because in Arvaniti & Ladd the wh-word with initial stress, ['pu], starts with a voiceless consonant, while the wh-word without initial stress, [apo'pu], starts with a vowel, an F0 contour with a rising onset is likely to start higher in the former case, as shown in Fig. 10. That is, a voiceless consonant perturbs the F0 contour of a syllable in two ways, raising the onset F0 very briefly, and 'truncating' an otherwise

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

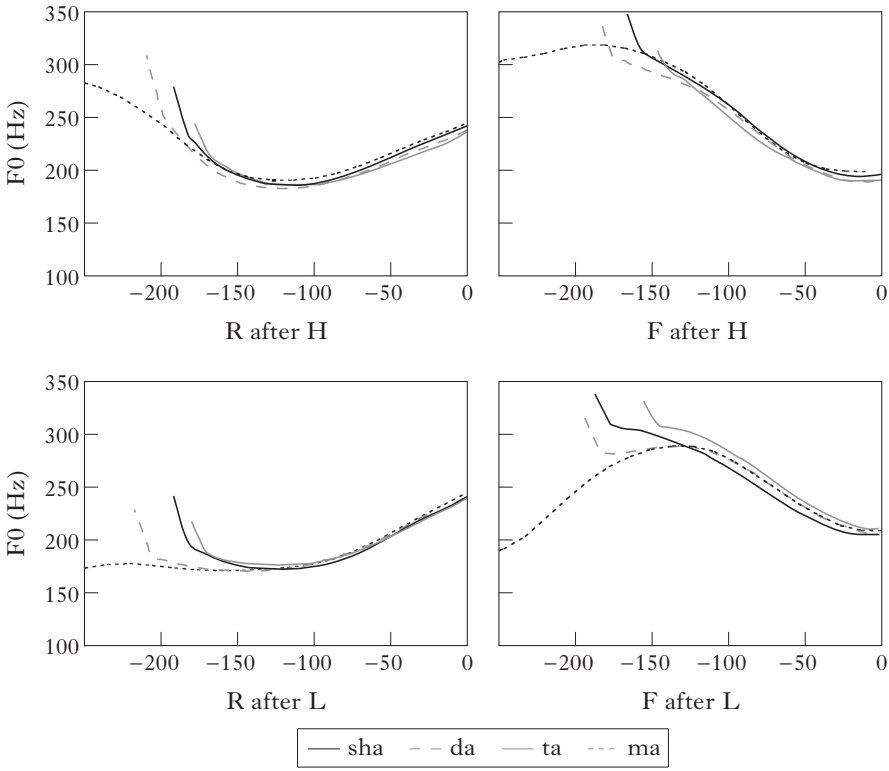


Figure 10

Effects of voiceless consonants on the F0 contours of Mandarin rising and falling tones produced after high and low tones. Each curve is an average across five repetitions, two carrier sentences and seven female speakers. All curves are aligned to the syllable offset. Data from Xu & Xu (2003).

continuous F0 movement, as can be clearly seen when compared to the F0 of a sonorant onset (Xu & Xu 2003). Such a 'truncation' mechanism has already been implemented in the PENTA trainers and is tested in Xu & Prom-on (2014).

5 Concluding remarks

We have presented an overview of PENTA as a framework for conceptually and computationally linking communicative meanings to fine-grained prosodic details, based on an articulatory-functional view of speech communication. In this framework a rich repertoire of communicative functions is simultaneously realised through an articulatory encoding process, so that all the details of the surface prosody can be traced back

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press

to their respective sources. As such, PENTA has addressed the major criteria advocated by Arvaniti & Ladd for a complete theory of intonation, namely ABSTRACTION, GENERALISATION, PREDICTION and ACCOUNTING FOR DETAIL.

Abstraction is addressed in PENTA by defining prosodic categories primarily in terms of communicative functions, while treating the underlying phonetic forms of the functional categories as a matter of empirical discovery. It is further achieved by the ability of the articulatory mechanisms simulated by qTA, with which an invariant (hence abstract) pitch target can generate an unlimited number of contextual variants (Xu & Prom-on 2014).

Generalisation is addressed in PENTA by treating the basic articulatory mechanisms of pitch production, as well as the core principle of encoding multiple layers of information in parallel, as universal, while allowing the phonetic details of the encoding schemes to be discovered through empirical studies.

Prediction is addressed in the PENTA approach at two levels. At the phonetic level, we have developed computational algorithms capable of learning function-specific pitch targets from natural speech and using the learned parametric representations to synthesise F0 contours that closely match those of natural utterances, either by the same speaker or by different speakers. At the functional level, prediction is addressed by always looking for the proper sources of the encoding schemes. Some of the sources are historical, and are thus responsible for language-specific variations; some are biological or bio-informational, hence are behind encoding properties that are not only universal among human languages, but are also shared with other animal communication systems (Xu, Kelly & Smillie 2013, Xu, Lee *et al*. 2013).

Accounting for detail is addressed in PENTA by developing analysis and modelling tools that are capable of processing many aspects of prosodic events, and by trying to link them to underlying sources in terms of either articulation or functional encoding. A substantial number of details in surface prosody have already been accounted for, including various alignment and scaling patterns, as discussed in this paper. More importantly, the quality of these accounts can be assessed in numerical terms through computational modelling, which makes it possible for even highly theoretical debates to be conducted with the help of detailed quantitative comparisons.

REFERENCES

Arvaniti, Amalia & D. Robert Ladd (2009). Greek wh-questions and the phonology of intonation. *Phonology* **26**. 43–74.

Bailly, Gérard & Bleicke Holm (2005). SFC: a trainable prosodic model. *Speech Communication* **46**. 348–364.

Beckman, Mary E. & Janet B. Pierrehumbert (1986). Intonational structure in Japanese and English. *Phonology Yearbook* **3**. 255–309.

Birkholz, Peter, Bernd J. Kroger & Christiane Neuschaefer-Rube (2011). Model-based reproduction of articulatory trajectories for consonant–vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing* **19**. 1422–1433.

Black, Alan & Andrew Hunt (1996). Generating F_0 contours from ToBI labels using linear regression. *Proceedings of the 4th International Conference on Spoken Language Processing* (*ICSLP 96*). Vol. 3. 1385–1388.

Bolinger, Dwight L. (1986). *Intonation and its parts: melody in spoken English*. London: Arnold.

Broe, Michael B. & Janet B. Pierrehumbert (eds.) (2000). *Papers in laboratory phonology V: acquisition and the lexicon*. Cambridge: Cambridge University Press.

Chen, Matthew Y. (2000). *Tone sandhi: patterns across Chinese dialects*. Cambridge: Cambridge University Press.

Chen, Yiya & Yi Xu (2006). Production of weak elements in speech: evidence from F_0 patterns of neutral tone in Standard Chinese. *Phonetica* **63**. 47–75.

Cooper, William E., Stephen J. Eady & Pamela R. Mueller (1985). Acoustical aspects of contrastive stress in question–answer contexts. *JASA* **77**. 2142–2156.

de Jong, Kenneth (2004). Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *JPh* **32**. 493–516.

Doupe, Allison J. & Patricia K. Kuhl (1999). Birdsong and human speech: common themes and mechanisms. *Annual Review of Neuroscience* **22**. 567–631.

Fujisaki, Hiroya (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In Peter F. MacNeilage (ed.) *The production of speech*. New York: Springer. 39–55.

Grice, Martine, D. Robert Ladd & Amalia Arvaniti (2000). On the place of phrase accents in intonational phonology. *Phonology* **17**. 143–185.

Gussenhoven, Carlos (2000). The boundary tones are coming: on the nonperipheral realization of boundary tones. In Broe & Pierrehumbert (2000). 132–151.

Gussenhoven, Carlos (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

Hart, Johan 't, René Collier & Antonie Cohen (1990). *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

Heldner, Mattias (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *JPh* **31**. 39–62.

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

Jun, Sun-Ah (ed.) (2005). *Prosodic typology: the phonology of intonation and phrasing*. Oxford: Oxford University Press.

Kochanski, Greg & Chilin Shih (2003). Prosody modeling with soft templates. *Speech Communication* **39**. 311–352.

Ladd, D. Robert (2008). *Intonational phonology*. 2nd edn. Cambridge: Cambridge University Press.

Lee, Albert, Yi Xu & Santitham Prom-on (2014). Modeling Japanese F0 contours using the PENTAtrainers and AMtrainer. *Proceedings of the 4th International Symposium on Tonal Aspects of Languages* (*TAL2014*). 164–167.

Liu, Fang & Yi Xu (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* **62**. 70–87.

Liu, Fang, Yi Xu, Santitham Prom-on & Alan Yu (2013). Morpheme-like prosodic functions: evidence from acoustic analysis and computational modelling. *Journal of Speech Sciences* **3**. 85–140.

Nick, Teresa A. (2014). Models of vocal learning in the songbird: historical frameworks and the stabilizing critic. *Developmental Neurobiology*. DOI:10.1002/dneu.22189.

O'Connor, J. D. & G. F. Arnold (1973). *Intonation of colloquial English: a practical handbook*. 2nd edn. London: Longman.

Peng, Shu-Hui (2000). Lexical versus 'phonological' representations of Mandarin sandhi tones. In Broe & Pierrehumbert (2000). 152–167.
Pierrehumbert, Janet B. (1980). *The phonology and phonetics of English intonation*. PhD dissertation, MIT.
Pierrehumbert, Janet B. (1981). Synthesizing intonation. *JASA* **70**. 985–995.
Pierrehumbert, Janet B. (2000). Tonal elements and their alignment. In Merle Horne (ed.) *Prosody: theory and experiment. Studies presented to Gösta Bruce*. Dordrecht: Kluwer. 11–36.
Pierrehumbert, Janet B. & Mary E. Beckman (1988). *Japanese tone structure*. Cambridge, Mass.: MIT Press.
Pierrehumbert, Janet B. & Julia Hirschberg (1990). The meaning of intonational contours in the interpretation of discourse. In Philip R. Cohen, Jerry Morgan & Martha E. Pollack (eds.) *Intentions in communication*. Cambridge, Mass.: MIT Press. 271–311.
Prom-on, Santitham, Peter Birkholz & Yi Xu (2013). Training an articulatory synthesizer with continuous acoustic data. *Proceedings of Interspeech 2013*. 349–353.
Prom-on, Santitham & Yi Xu (2012). PENTATrainer2: a hypothesis-driven prosody modeling tool. In Antonis Botinis (ed.) *Proceedings of the 5th IESL Conference on Experimental Linguistics,* Athens, Greece. 93–100.
Prom-on, Santitham, Yi Xu & Bundit Thipakorn (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *JASA* **125**. 405–424.
Raidt, S., G. Bailly, B. Holm & H. Mixdorff (2004). Automatic generation of prosody: comparing two superpositional systems. In Bernard Bel & Isabelle Marlien (eds.) *Speech prosody 2004*. Nara, Japan. Available (October 2015) at http://www.isca-speech.org/archive/sp2004. 417–420.
Saltzman, Elliot & Kevin G. Munhall (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology* **1**. 333–382.
Sun, Xuejing (2002). *The determination, analysis, and synthesis of fundamental frequency*. PhD dissertation, Northwestern University.
Taylor, Paul (2000). Analysis and synthesis of intonation using the Tilt model. *JASA* **107**. 1697–1714.
Wang, Bei & Yi Xu (2011). Differential prosodic encoding of topic and focus in sentence-initial position in Mandarin Chinese. *JPh* **39**. 595–611.
Xu, Ching X. & Yi Xu (2003). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association* **33**. 165–181.
Xu, Ching X., Yi Xu & Li-Shi Luo (1999). A pitch target approximation model for F_0 contours in Mandarin. In John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville & Ashlee C. Bailey (eds.) *Proceedings of the 14th International Congress of Phonetic Sciences*. Berkeley: University of California. 2359–2362.
Xu, Yi (1997). Contextual tonal variations in Mandarin. *JPh* **25**. 61–83.
Xu, Yi (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication* **46**. 220–251.
Xu, Yi (2011a). Speech prosody: a methodological review. *Journal of Speech Sciences* **1**. 85–115.
Xu, Yi (2011b). Post-focus compression: cross-linguistic distribution and historical origin. In Wai-Sum Lee & Eric Zee (eds.) *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong 2011*. Hong Kong: University of Hong Kong. 152–155.
Xu, Yi, Szu-Wei Chen & Bei Wang (2012). Prosodic focus with and without post-focus compression: a typological divide within the same language family? *The Linguistic Review* **29**. 131–147.

Xu, Yi, Andrew Kelly & Cameron Smillie (2013). Emotional expressions as communicative signals. In Sylvie Hancil & Daniel Hirst (eds.) *Prosody and iconicity*. Amsterdam & Philadelphia: Benjamins. 33–59.

Xu, Yi, Albert Lee, Wing-Li Wu, Xuan Liu & Peter Birkholz (2013). Human vocal attractiveness as signaled by body size projection. *PLoS ONE* 8. e62397. Available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0062397.

Xu, Yi & Fang Liu (2006). Tonal alignment, syllable structure and coarticulation: toward an integrated model. *Rivista di Linguistica* **18**. 125–159.

Xu, Yi & Fang Liu (2012). Intrinsic coherence of prosodic and segmental aspects of speech. In Oliver Niebuhr (ed.) *Understanding prosody: the role of context, function and communication*. Berlin & Boston: de Gruyter. 1–26.

Xu, Yi & Santitham Prom-on (2010–14). PENTAtrainer1: a Praat script for extracting pitch targets from individual sound files. Available (October 2015) at http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1.

Xu, Yi & Santitham Prom-on (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, Yi & Q. Emily Wang (2001). Pitch targets and their realization: evidence from Mandarin Chinese. *Speech Communication* **33**. 319–337.

Xu, Yi & Ching X. Xu (2005). Phonetic realization of focus in English declarative intonation. *JPh* **33**. 159–197.

https://doi.org/10.1017/S0952675715000299 Published online by Cambridge University Press