To appear in Phonology

Explaining the PENTA model:

A reply to Arvaniti & Ladd (2009)

Yi Xu

Albert Lee

University College London

Santitham Prom-on

King Mongkut's University of Technology Thonburi

Fang Liu

University College London

Abstract

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 and 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 and Ladd (2009), showing how it is possible to predict the prosodic shapes of an utterance based on the lexical and post-lexical meanings it conveys.

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, we believed, had not received sufficient attention, namely, communicative functions and articulatory mechanisms (2005). The goal was to develop a framework that would explain how speech prosody works as a system of communication. More specifically, the framework 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 that 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 et al. 2013, Liu & Xu 2005, Prom-on et al. 2009, Wang & Xu 2011, Xu & Liu 2012, Xu & Prom-on 2014).

PENTA has received much scrutiny since its proposal, and one of the most comprehensive critiques is offered by Arvaniti and Ladd (2009; henceforth A&L). A&L have contrasted PENTA with the Autosegmental-Metrical (AM) theory of prosody (Beckman & Pierrehumbert 1986, Gussenhoven 2004, Ladd 2008, Pierrehumbert 1980, Pierrehumbert & Beckman 1988), and argued that it is inadequate to explain the prosody of Greek wh-questions examined in the study. Such a direct theoretical comparison is welcome, as it gives us 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 specific criticisms by A&L. Finally, we will offer hypothetical interpretations of the prosody of Greek wh-questions based on data presented by A&L, 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 punctuations, 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 communicative function. For example, for the pitch track shown in Figure 1, it is hard to determine what the relevant prosodic units are: F_0 peaks and valleys, turning points, size of the F_0 movements, temporal scope of a continuous movement, or all of them, or none of them? 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.

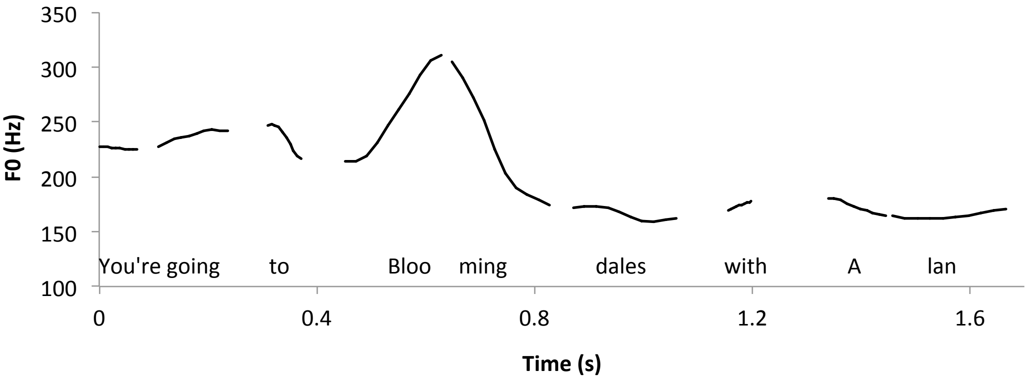


Figure 1. F_0 track of "You're going to *Bloomingdales* with Alan" by a female American English speaker, with focus on "Bloomingdales". Data from Liu et al. (2013).

The strategy adopted by AM 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 AM framework of prosody, which encompasses a rich inventory of phonological primitives that form the intonation systems of English (Ladd 2008, Pierrehumbert 1980) as well as many other languages (Jun 2005). Overall, although there are variations in theoretical details and methodological approaches among studies in the AM tradition, the central question in this approach remains the same: What does the phonology of speech prosody look like? (Gussenhoven, 2004, Ladd, 2008, Pierrehumbert 1980)

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: a) What are the meanings that are conveyed by prosody? b) How does prosody encode these meanings in a way that allows easy decoding? In our search for answers, we took a bootstrapping strategy of always keeping one side of the function-form link relatively unambiguous when 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 production in connected speech, as summarized in Xu (2005, 2011). These studies established that even syllable-bound lexical tones do not show stable F_0 properties in connected speech, but exhibit extensive surface variability with tonal contexts (which is contrary to A&L's claim [p. 65] that PENTA assumes stable syllable-by-syllable specification of F_0 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 F_0 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, and 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. To PENTA, 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 already a standard principle in phonology, the special challenge of prosody, as mentioned above, has motivated us to insist on the primacy of function in the function-form relation, especially in case of uncertainty. For example, the long-standing AM debate over whether LH* and H* are distinct phonological categories in English prosody (Ladd 2008) is to PENTA a non-issue, 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., are attributed to obligatory articulatory processes rather than to phonology.

One thing that PENTA does share with AM 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 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 (Xu 2005). PENTA assumes that similar arbitrary rules also exist in prosody, and it is based on this assumption that 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 which will be illustrated in 3.2.

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 first block from the left 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 stack 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 specific encoding schemes. What we have always assumed, though not necessarily made fully explicit each time, is that some of the encoding schemes are highly stylised and language specific, while others are more gradient and universal. The third block in Figure 2 represents the articulatory parameters that are linked to the encoding schemes. These parameters in turn control the target approximation process represented by the fourth block. It is this mechanical process that directly generates surface acoustics, including F_0,[1] through the mechanism of *target approximation* (TA) as represented by the fourth block from the left in Figure 2.

[1] As postulated in Xu and Liu (2006, 2012) and recently tested in Prom-on, Birkholz & Xu (2013), the notion of underlying targets applies not only to F_0, 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.

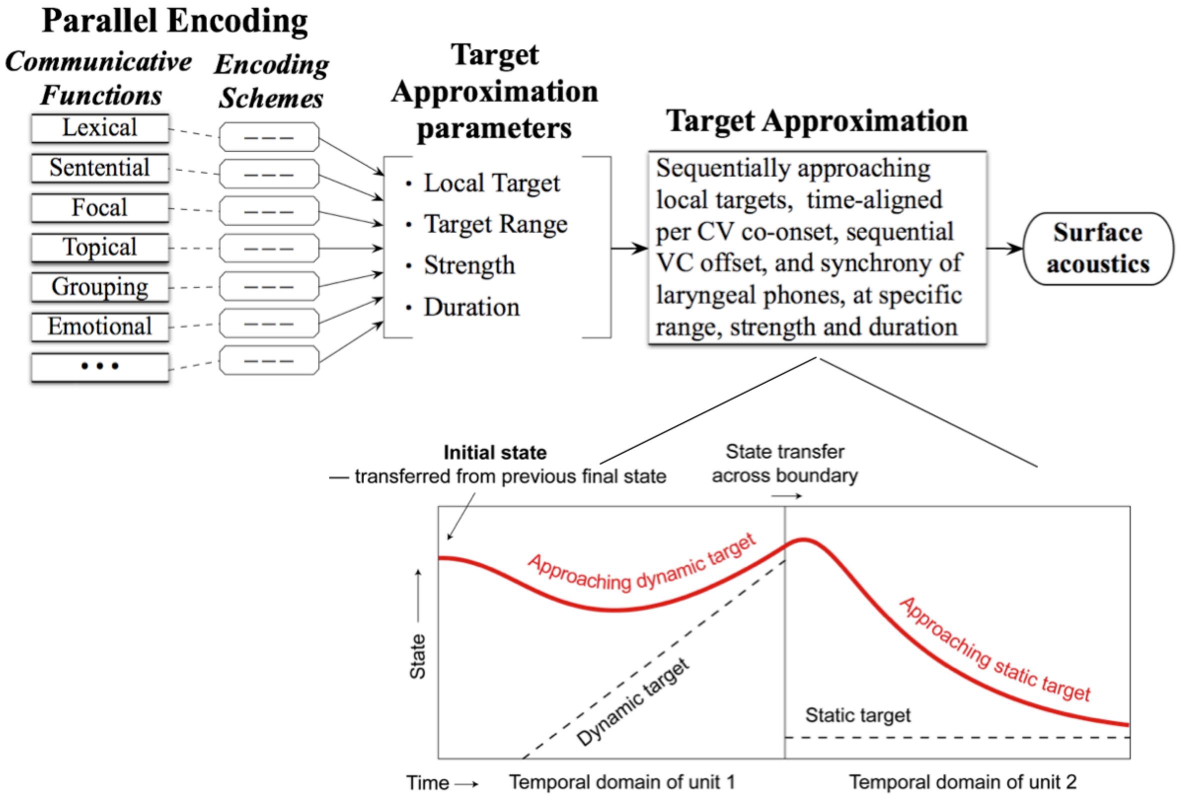

Parallel Encoding
Communicative Functions
Encoding Schemes
Lexical
Sentential
Focal
Topical
Grouping
Emotional
• • •
Target Approximation parameters
• Local Target
• Target Range
• Strength
• Duration
Target Approximation
Sequentially approaching local targets, time-aligned per CV co-onset, sequential VC offset, and synchrony of laryngeal phones, at specific range, strength and duration
Surface acoustics
Initial state
— transferred from previous final state
State transfer across boundary
Approaching dynamic target
Approaching static target
Dynamic target
Static target
State
Time
Temporal domain of unit 1
Temporal domain of unit 2

Figure 2. Upper panel: A schematic sketch of the PENTA model. Lower panel: The target approximation component of PENTA, which is an articulation process (Xu 2005, Xu & Liu 2012, Xu & Wang 2001).

The TA model, as depicted graphically in the lower panel of Figure 2, assumes that each syllable is assigned an underlying target that has not only a height (or position), but also a slope specification. The surface F_0 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 F_0 turning point, as depicted in the lower panel.

There are many implications of PENTA that may not be immediately obvious just from the descriptions given above, and this has often led to confusions about the model. Thus, it would be helpful to lay out some of the most critical implications of PENTA, first in a brief list, as shown below, followed by further elaborations in the subsequent sections.

(1) Syllable-sized pitch targets are the prosodic primitives in PENTA, and as such they bear the closest resemblance to tones in AM theory. They differ from the AM-tones in that their link to surface F_0 trajectories is via syllable-synchronised sequential target approximation. In contrast, linear or sagging interpolation between specified targets proposed by Pierrehumbert (1980) is the assumed mechanism in AM theory, as is made clear in A&L. For PENTA, as shown in Figure 2, because they do not directly correspond to observable features such as turning points, elbows or plateaus, all targets are virtual.

(2) 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.

(3) For each syllable, a unique target is assigned as a result of the interaction of all the communicative functions involved (as indicated in Figure 2 by the single arrow between the *Target Approximation parameters* block and the *Target Approximation* block versus the multiple arrows between the *Parallel Encoding* block and the *Target Approximation parameters* block). Thus the encoding schemes of all the involved functions jointly determine a unique target of each syllable for a particular phonetic dimension. This *integrated* target therefore carries the information of all the encoded functions.[2]

(4) In contrast to its explicit assumption about articulatory mechanisms, PENTA has no explicit stipulation on a pre-defined inventory of communicative functions or their encoding schemes for any language. Rather, it assumes that encoding schemes, be they language specific or universal, categorical or gradient, have to be established experimentally by directly controlling communicative functions.

(5) 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 through both articulatory mechanisms that are universal, and encoding schemes that are either universal or language-specific.

(6) 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 AM theory can be reinterpreted, from the PENTA perspective, as being morpho-phonological 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 pre-defined 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 pre-defined in other theories, such as H, H*, L, L* in AM theory.

2.3 Recent new conceptual development

More recently there has been a further development in our conceptualisation of the encoding schemes in PENTA (Liu et al. 2013). This was driven by our 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, and these components are meaningless by themselves, but act jointly to mark both intra- and inter-functional contrasts. Second, similar to lexical morphemes, an encoding scheme for a prosodic function may have allomorph-like variants

[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 F_0 contours, which are mathematically combined at the final stage of the model computation to form the ultimate surface F_0 contours.

whose occurrence is conditioned by factors like location in sentence and interaction with other prosodic functions. Finally, similar to lexical morphemes, these encoding schemes are language-specific and their patterns likely have historical origins.[3] These prosodic encoding schemes differ from lexical morphemes in that they contrast prosodic functions that carry post-lexical meanings. It is therefore appropriate to refer to them collectively as prosodic morphemes or *prosophemes*.

One of the clearest examples of prosophemes 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 (see review by Xu et al. 2012). Also, in many languages, focus is realised not only with prosodic patterns of the focused unit itself, but also with *post-focus compression* (PFC) of pitch and intensity (see review by Xu, Chen & Wang 2012). Furthermore, PFC has recently been found to be absent in many other languages (Xu et al. 2012). It is hypothesised that PFC 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 with likely historical etymologies — very similar to lexical morphemes.

Another example is that, in American English, the underlying pitch target of a stressed syllable varies depending on whether the syllable is word final or non-final, whether the word is focused or not, and whether the sentence is a statement or yes-no question (Liu et al. 2013), as can be seen in Figure 3. Figure 3 also shows that the F_0 of the post-focus syllables varies markedly depending on whether the sentence is a statement or question. In particular, post-focus F_0 in a question is raised well above the reference level, i.e., the pre-focus F_0. This pattern, however, is absent in Mandarin (Liu et al. 2013), as can be seen in Figure 4. Such a cross-linguistic typological difference is again similar to the behavior of lexical morphemes, although more research is needed to further explore this phenomenon.

The prosopheme notion is an alternative to the tonal morpheme proposed by Pierrehumbert and Hirschberg (1990). As discussed in detail in Liu et al. (2013), many of the morpheme-like meanings proposed by Pierrehumbert and 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 AM 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 F_0 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 (Figure 3). Thus this extra raising is morpho-phonological, i.e., being part of a prosopheme, rather than being a phonetic implementation rule.

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

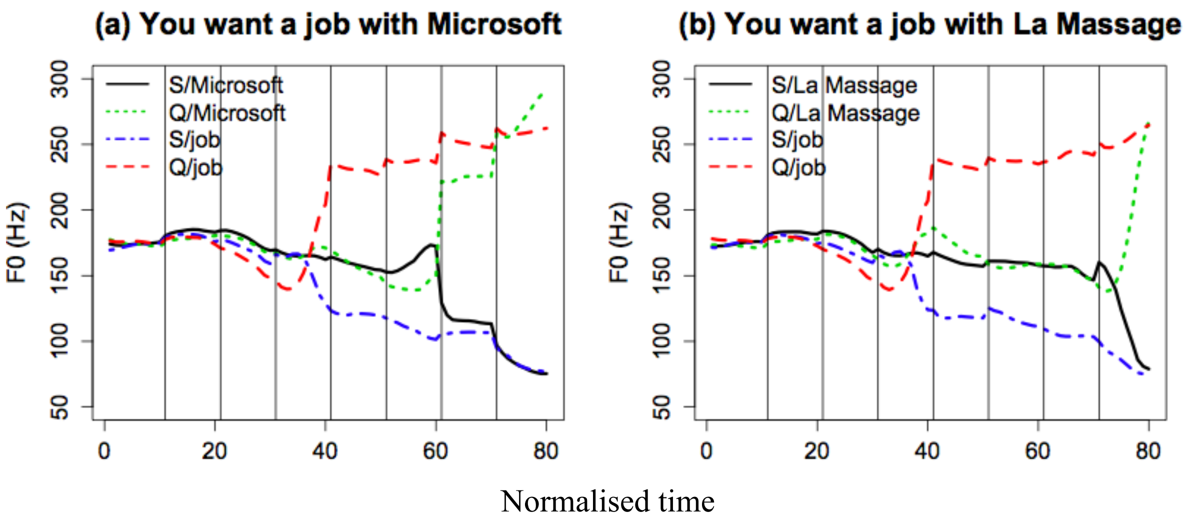


Figure 3. Mean F_0 contours of statements (S) and questions (Q) in American English. The word after "/" is focused. Data from Liu et al. (2013).

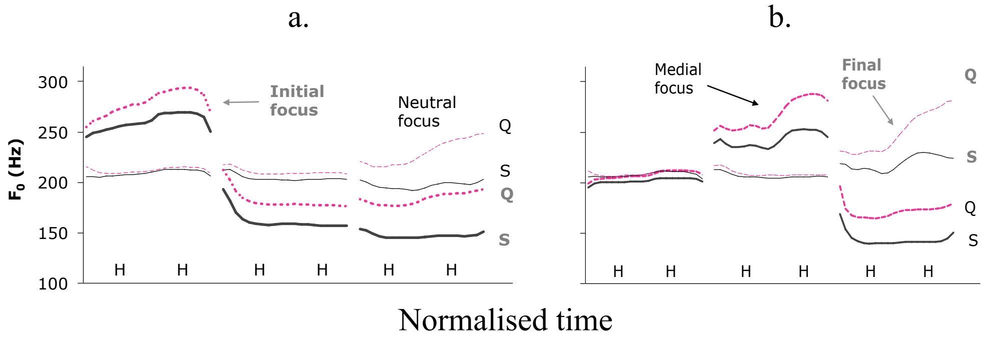


Figure 4. Mean F_0 contours of the Mandarin sentence Zhāng Wēi dānxīn Xiāo Yīng kāichē fāyūn 张威担心肖英开车发晕 [Zhang Wei is concerned that Xiao Ying may get dizzy when driving] spoken as either a statement or a question. On the left, either focus is on the sentence-initial word (thick lines), or there is no narrow focus (thin lines). On the right, focus is either sentence medial (thick lines) or sentence final (thin lines). The black solid lines represent statements, and the pink dashed lines represent questions. Data from Liu and Xu (2005).

2.4 A quantitative realisation

Like most other theoretical intonation models ('t Hart et al. 1990, Bolinger 1986, O'Connor & Arnold 1973, Pierrehumbert 1980), PENTA was qualitative at the time of its proposal (Xu 2005). As such it could be applied 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 first 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 as 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 F_0 of each syllable is represented by a third-order critically damped linear system driven by a pitch target, as shown in the following equation,

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

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 F_0 dynamic state and pitch target of the current syllable. As such they are not free parameters. Parameter λ represents the strength of the F_0 movement toward the target. qTA realises the state transfer between adjacent syllables by taking the final F_0 state of the preceding syllable in terms of its final F_0, $f_0(0)$, velocity, $f_0'(0)$ and acceleration, $f_0''(0)$ as the initial F_0 dynamic state of the current syllable. With this initial state the three transient coefficients are computed with the following formulae,

$$c_1 = f_0(0) - b \tag{2}$$

$$c_2 = f_0'(0) + c_1\lambda - m \tag{3}$$

$$c_3 = \left(f_0''(0) + 2c_2\lambda - c_1\lambda^2\right)/2 \tag{4}$$

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 raising and lowering of pitch targets relative to the speaker's average F_0. λ indicates how rapidly a pitch target is approached, with higher values representing faster target approximation. 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 *ex post facto*. Despite the similarities, however, at least three key features remain unique to qTA after a close examination: a) unitary dynamic targets (which are different from contour targets as in Stem-ML (Kochanski & Shih 2003) or SFC (Bailly & Holm 2005) models), b) 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), c) high-order state transfer across target approximation movements, a feature not found in any other model except VocalTractLab, which adopted the same idea and made the transfer order even higher (Birkholz, Kroger & Neuschaefer-Rube 2011).

2.5 Why pitch target for every syllable?

One of the most questioned aspects of PENTA is its assumption of 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 over-fitting for languages that are not lexically tonal. In English and Greek, for example, many syllables appear unspecified for pitch because of their high F_0 variability, absence of prominent peaks or valleys, and lack of stress. It therefore seems natural to assume, as does AM theory, that 'not every syllable has to have a specification for pitch' (A&L p. 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 Figure 2. That is, the F_0 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 F_0 contour for a syllable besides 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 is done in the Fujisaki model. There are two reasons why we choose not to do so. The first is that it is our assumption that the syllable, as a basic coarticulatory unit, is produced with all its underlying targets fully specified, be it 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, Xu & Liu 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 F_0 contours to be generated separately and then added to the syllable. The second 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 and 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 weak strength, which is consistent with its weak stress status. As found in both acoustic analysis (Chen & Xu 2006, Xu & Xu 2005) and computational modelling (Liu et al. 2013, Xu & Prom-on 2014), such weak 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 account for, at least hypothetically, the alignment patterns in Greek wh-questions reported by A&L.

As a further support, there is evidence that computational models with F_0 specifications for every syllable generate synthetic prosody with better numerical and perceptual quality than those that have non-syllabic pitch specifications (Raidt et al. 2004, Sun 2002). Sun (2002), in particular, found that the three-target model (Black & Hunt 1996), which simply uses three F_0 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 F_0 peaks, when both models were trained on the same corpus.

Finally, in terms of economy of representation, each target per syllable may not be as uneconomical as it may appear. 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 versus 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 temporal alignment of onset and offset of prosodic units relative to segments, as is obligatory in models that assume flexible timing (Fujisaki 1983, Pierrehumbert 1981).[4] As

[4] On this point 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 (as in the Fujisaki

will be presented in greater detail in Section 3, only a small number of target parameters are needed to represent lexical tone, lexical stress, focus and modality in English, Mandarin and Thai (Prom-on et al. 2009, Prom-on & Xu 2012). 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 F_0 contours.

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 the Introduction, which is to develop a model that can explain exactly how speech works as a communication system. Based on this tenet, 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 main task in the PENTA approach is therefore to identify the encoding schemes of various communicative functions. Our 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 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 the articulatory mechanisms, some of which are already built into PENTA. For example, articulatory inertia makes it impossible for F_0 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 constraint. For example, the cross-linguistic distribution of PFC found in recent research, as discussed in 2.3, has led to the Nostratic origin of PFC hypothesis, which makes strong predictions about the existence of PFC 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 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 indispensible parts of the speech communication process.

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

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.

as mentioned above, it is imperative that the assumed mental representation is operable. What this means is that, first, 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 our effort to develop a computational realization 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 based on specific articulatory mechanisms that can be simulated with a computational model. For PENTA, such a model is qTA, as introduced in 2.4. 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 the proposal of qTA, we have been developing computational tools that enable its conceptual exploration and quantitative testing. So far four tools have been developed. qTA_demo1 (http://www.phon.ucl.ac.uk/home/yi/qTA/), which was mentioned by A&L, and qTA_demo2 (http://www.phon.ucl.ac.uk/home/yi/qTA_demo2/). Both are web-based interactive Java programs that demonstrate how the qTA model works. 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 Figure 7 to be discussed later).

The other two tools, PENTAtrainer1 (Xu & Prom-On 2010-2014) and PENTAtrainer2 (Xu & Prom-on 2014), are data-driven modelling programs (http://www.phon.ucl.ac.uk/home/yi/PENTAtrainers/). 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 F_0 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 F_0 generation is assessed by three means: a) root mean squared errors (RMSE), which measures the discrepancy of the synthetic contours from the original contours in terms of point-by-point height difference, b) Pearson's *r*, which assesses how closely the overall shape of the synthetic contours correlates with that of the original contours, and c) perceptual evaluation in terms of category identification (e.g., tone, focus, etc.) and naturalness.

Critically, both PENTAtrainers allow predictive synthesis of F_0 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 equation 1). The parameter set that achieves the best fit to the F_0 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 Figure 5, where the short dashed lines are the learned targets. The F_0 contours generated with these learned targets (solid curves) seem to fit the original F_0 contours (dotted curves) 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.

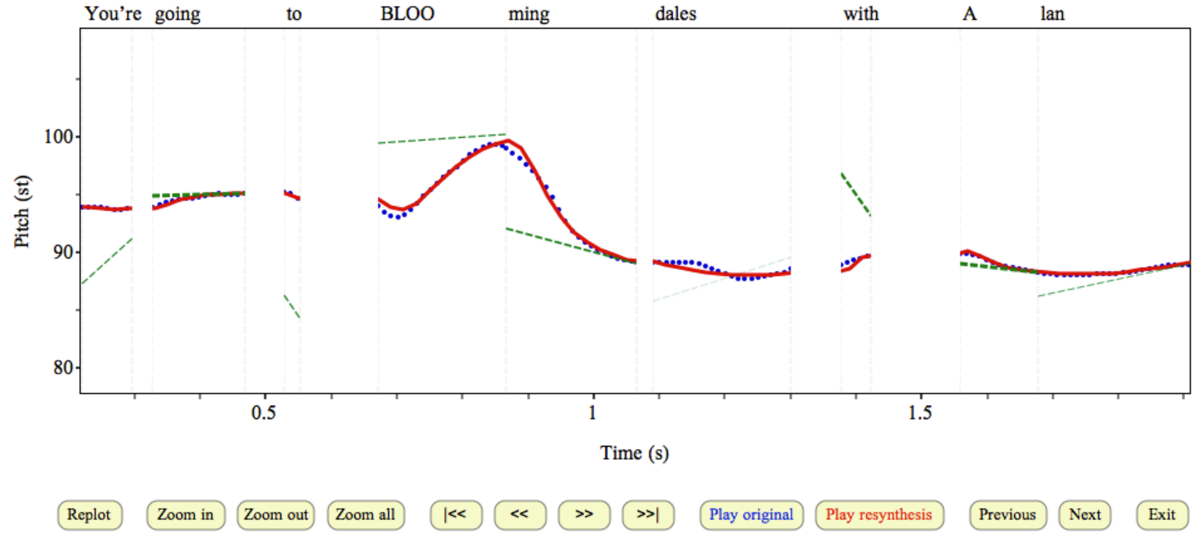
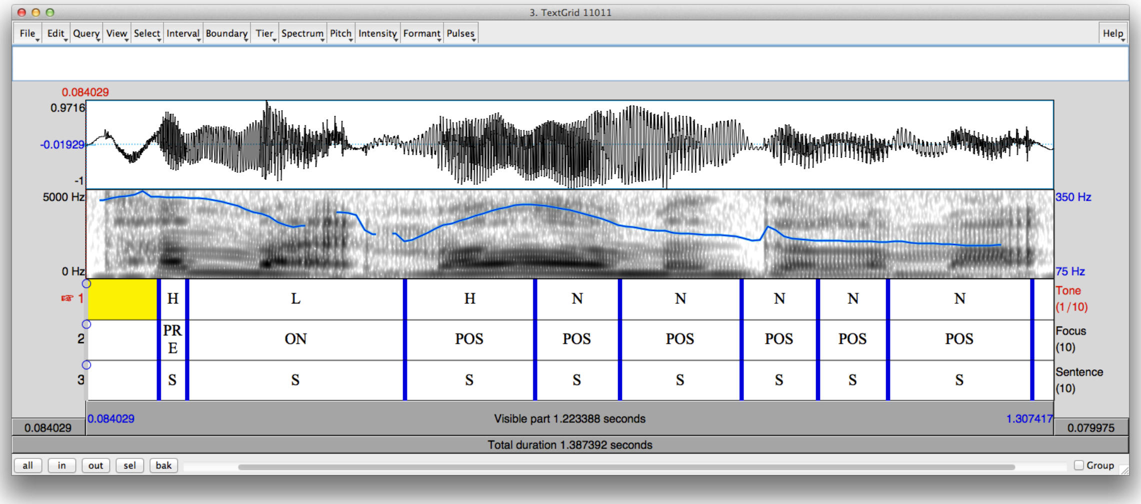


Figure 5. Original (blue dotted) vs. resynthesised (red solid) F_0 contours of the English utterance "You're going to Bloomingdales with Alan" shown in Figure 1, by PENTAtrainer1. [X-axis: Time (seconds), Y-axis: F_0 in semitones.]

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 ecologically unrealistic as a learning algorithm. 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 F_0 and synthetic contours (bottom), where the latter is generated with categorical target parameters learned from an entire corpus.



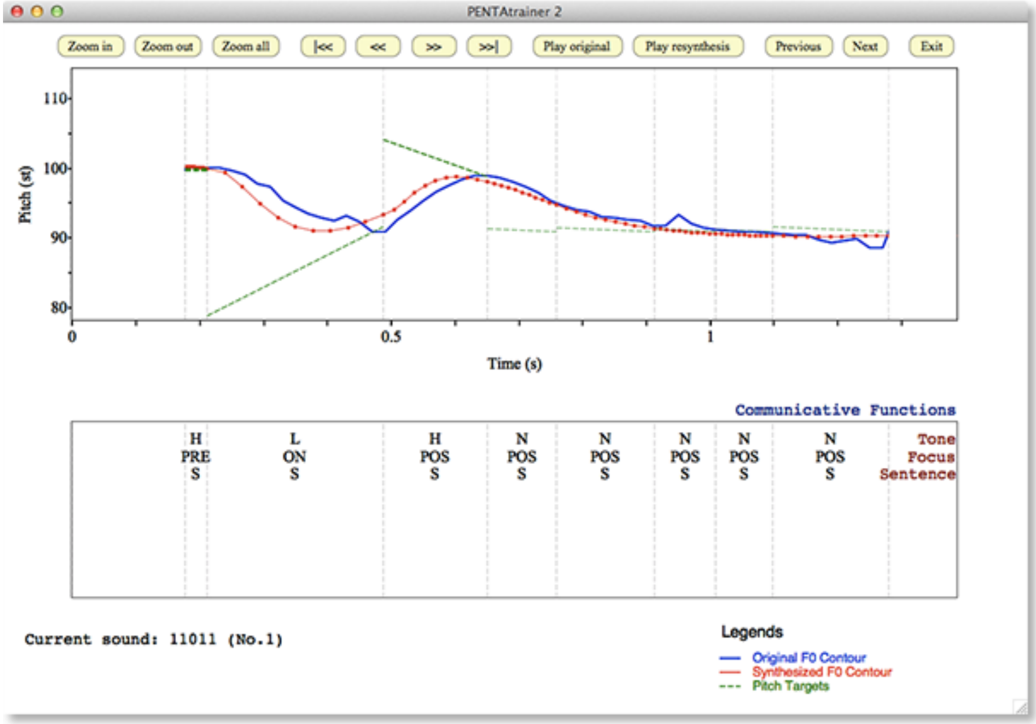


Figure 6. Snapshots of PENTAtrainer2 interfaces (www.phon.ucl.ac.uk/home/yi/PENTAtrainer2/). The annotation interface (top) allows users to mark temporal scope of functional units. In this example (Mandarin sentence "tā MǍI māma men de la ma" [Did he BUY what mother has?], with focus on mai3), the annotated functions are lexical tone, focus and sentence type (modality). All the boundaries 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), synthetic (dotted) and natural (solid) F_0 contours. It also allows users to play the utterance with either synthetic or natural prosody (Prom-on & Xu 2012).

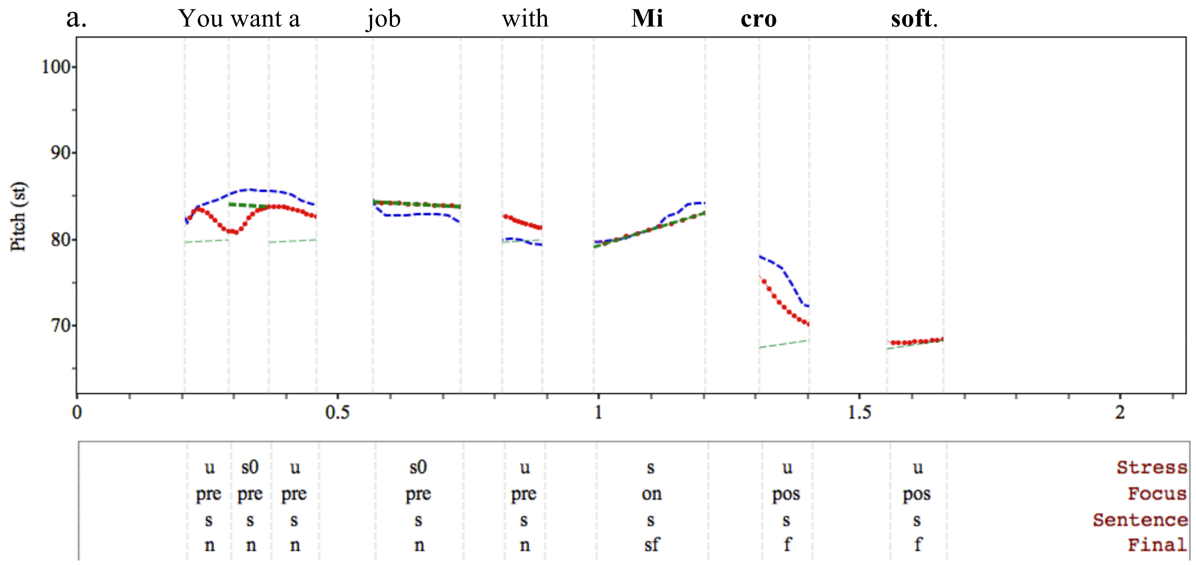
In Xu & Prom-on (2014), we achieved good overall numerical results with PENTAtrainer2 for English (the same dataset tested with PENTAtrainer1 in Liu et al. 2013), Mandarin and Thai. In Prom-on et al. (2009), which applied categorisation by averaging, the perceptual identification rates for tone in Mandarin 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 F_0 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 b, m, λ values each) 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. Here the number of function-specific parameters roughly equals the number of parameters per target times the number of simulated functions times the number of function-internal categories minus non-existing category combinations (function-specific parameters = parameters per target × simulated functions × functional-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., 5 parameters per Standard Chinese tone in the Fujisaki model (Fujisaki 1983) and 4 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 has allowed 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 F_0 of an American English utterance, pitch targets learned by PENTAtrainer2, and synthetic F_0 contours generated with the learned targets. The sentences were spoken with either sentence-medial or sentence-final focus, and as either statements or 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:

1. Focus is characterised by a robust post-focus pitch range shift, with the direction of the shift dependent on modality: downward in a statement (a, b), but upward in a question (c, d). The resulting post-focus plateaus correspond to the L- and H- phrase accents in AM, but from the PENTA perspective, they are allomorphic components of the focus and modality encoding schemes (or prosophemes), rather than autonomous prosodic units in their own right.
2. 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 b, d). For on-focus, non-final stressed syllables, the target slope rises in both statements and questions, at least for this speaker (*Mi-* in a, c).
3. In both statements and questions, targets are higher in stressed syllables than in unstressed syllables; but the differences are much smaller in questions.



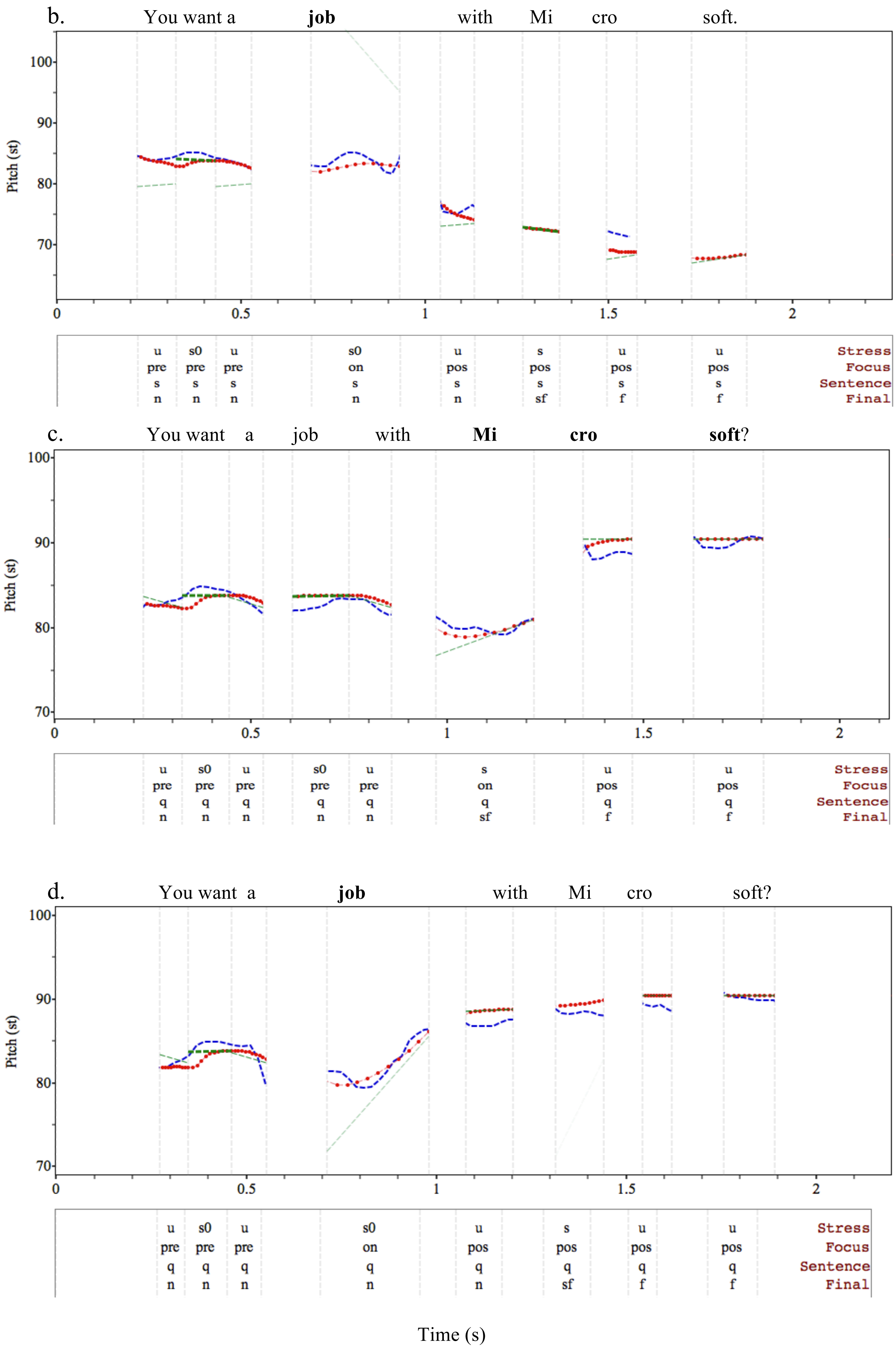

b.
You want a
job
with
Mi
cro
soft.
Pitch (st)
100
90
80
70
0
0.5
1
1.5
2
u pre s n
s0 pre s n
u pre s n
s0 on s n
u pos s n
s pos s sf
u pos s f
u pos s f
Stress
Focus
Sentence
Final
c.
You want
a
job
with
Mi
cro
soft?
Pitch (st)
100
90
80
70
0
0.5
1
1.5
2
u pre q n
s0 pre q n
u pre q n
s0 pre q n
u pre q n
s on q sf
u pos q f
u pos q f
Stress
Focus
Sentence
Final
d.
You want
a
job
with
Mi
cro
soft?
Pitch (st)
100
90
80
70
0
0.5
1
1.5
2
u pre q n
s0 pre q n
u pre q n
s0 on q n
u pos q n
s pos q sf
u pos q f
u pos q f
Stress
Focus
Sentence
Final
Time (s)

Figure 7. Original (blue dashed) and synthetic (red dotted) F_0 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 (green dashed lines) learned by PENTAtrainer2 based on the functional annotations shown at the bottom of each graph. For stress, u = unstressed, S = non-final stressed, and s0 = word-final stressed. For syllable position (labeled as Final), n = non-final, sf = semifinal, and f = sentence final. All the graphs are screenshots of the demo window of the Synthesis tool (*synthesize.praat*) in the PENTAtrainer2 package. Data from Xu and Prom-on (2014).

Compared to the F_0 contours in Figure 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. Also unlike in English, Mandarin tones do not change the direction of their target slopes from statements to questions, presumably due to the lexical tonal constraint. These cross-linguistic differences in the encoding schemes of similar prosodic functions show that they are highly language-specific, and their exact forms cannot be predicted solely on functional grounds.

Note also that the match between the synthetic and original F_0 contours in Figure 7 is not nearly as good as that in Figure 5. This is partly because here the synthesis is predictive, based on categorical parameters learned from all the utterances by a speaker in a corpus, as opposed to resynthesis in Figure 5 (by PENTAtrainer1), but partly also because there is still room for further adjustments in our 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 F_0 in *-crosoft* does not show final upstep in Figure 7d. Thus, even if the major characteristics of the encoding schemes have been 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 Figure 6. These targets are abstract, as each of them is specified by only three parameters, but can correspond to 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 1 displays these properties, and shows which of them are shared by symbolic representations. As can be seen, only abstractness is unquestionably shared by both types of representations. Although it is possible to obtain AM-style representations in a data-driven manner (Lee, Xu & Prom-on 2014), its predictive power is yet unknown.

Table 1. Comparison of PENTA-based parametric and AM-style symbolic representations.

Properties	**Parametric**	**Symbolic**

Functionally defined	√	
Abstract (free of redundant and variant surface details)	√	√
Model-based (with mathematically defined articulatory mechanisms)	√	
Gradient (allowing for individual and dialectal variations)	√	
Data-driven (trainable, learnable)	√	?

Model-based parametric representations may also offer a solution to a well-known puzzle in phonology, namely, tone sandhi (Chen 2000). For example, the Mandarin Tone 3 is changed to T2 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 by the strength of the boundary between the two syllables. Such functional 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 do not need to be identical, since the functional combinations are not the same. As found in Xu and Prom-on (2014), 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 (Peng 2000, Xu 1997). 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 this line 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 PENTAtrainers 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., 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 PENTA interpretations of Greek wh-question prosody

Because the present paper is prompted by A&L's criticism of PENTA based on their Greek data, it is necessary for us to offer a PENTA-based interpretation of what A&L have reported about Greek wh-question prosody. But 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 by Figure 7, which displays functional annotations (bottom tiers), corresponding to hypothetical underlying pitch targets (dashed lines) and qTA-simulated F_0 contours of two sentences from A&L, 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) (syllable 1 in Figure 7a and syllables 1-3 in Figure 7b) but lowers the pitch targets of all subsequent syllables. The raised on-focus pitch targets result in an early F_0 peak, but the slope of the on-ramp of the peak depends on the lexical stress of the syllable: sharper if it is stressed (left graph), but shallower if it is unstressed (right graph). The lowered post-focus pitch targets result in an F_0 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 (Figure 7a), slower if it is unstressed (Figure 7b). The post-focus lowering also results in a low plateau after a post-focus stressed syllable (left graph). 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 (Liu et al. 2013, Prom-on et al. 2009, Xu & Prom-on 2014, Xu & Xu 2005), because there is no sufficient information about stress-related target height available from the data reported by A&L. The sentence-final rise, which involves a shallow rising target if the final syllable is unstressed (left), or a steep rising target if the final syllable is stressed (right), 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 progressive rise throughout the sentence, if Greek is similar to Mandarin: Liu & Xu 2005); lexical stress shows (hypothetically) consistent higher versus lower pitch targets. Each of these functional equivalences is shared by all the wh-question sentences presented by A&L, regardless of their length or lexical composition.

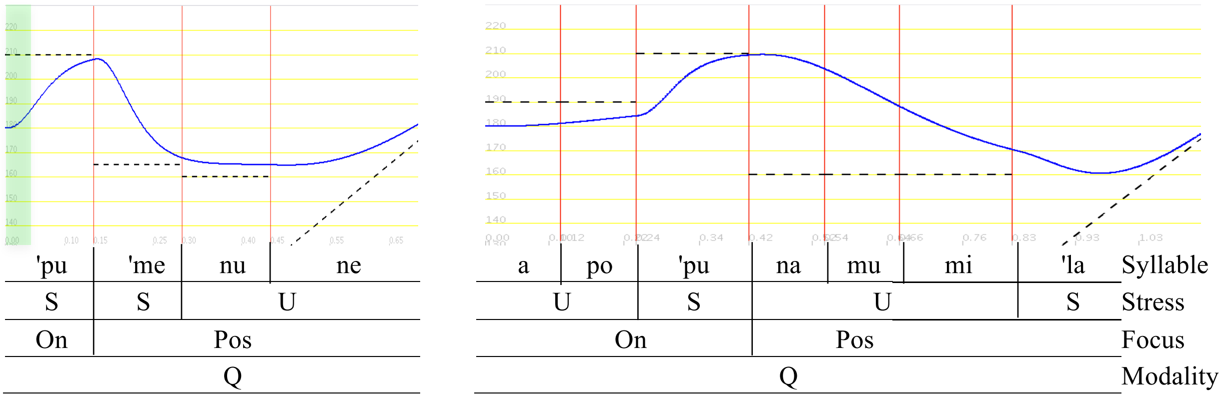


Figure 7. F_0 contours of ['pu 'menune] 'Where are they staying?' (left) and [apo'pu na mu mi'la] 'Where could s/he be talking to me from?' (right), resulting from qTA realisation of underlying pitch targets (dashed lines), which are hypothetically obtained from the functions annotated in the bottom tiers. Simulated using qTA_demo1 (http://www.phon.ucl.ac.uk/home/yi/qTA/). The translucent vertical bar at the left edge of the left graph is an illustration of the "truncation" effect of a voiceless consonant.

The pitch targets, represented by the dashed lines, which are purely hypothetical in the figure, can be obtained by applying PENTAtrainer1 or PENTAtrainer2 to the real data. The annotation tiers below the F_0 tracks illustrate PENTA-style functional annotations. Tier 1 on the top marks syllables and their boundaries; tier 2 marks lexical stress (S = stressed, U =

unstressed); tier 3 marks focus regions (On = on-focus, Pos = post-focus); and tier 4 marks sentence modality (Q = question).

In addition to the global patterns, Figure 7 also shows micro patterns related to alignment, scaling, etc., which are of major concern in A&L. 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 A&L the local variations are described in terms of alignment and scaling of F_0 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 F_0 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 F_0 adjustment mechanism in its own right. As is shown in Figure 7 and will be shown further, 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 F_0 peak

NH (nuclear H) measures the location of the early F_0 peak in a wh-question. A&L 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 in between, but there is no further variation in the 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 next section for definition of L1), the peak cooccurred with the onset consonant of the postnuclear syllable." (p. 58) A&L 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 Figure 7. In the sentence on the left, 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 F_0 peak very close to the syllable boundary. In contrast, in the sentence on the right, the first post-focus syllable is unstressed, thus has weak target strength. As a result, it takes longer for the on-focus rising momentum to be reversed, leading to an F_0 peak that is aligned more 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 A&L can be accounted for by PENTA using qTA simulation without any explicit specification of F_0 peak alignment or assumption of tonal crowding.

4.2.2 Alignment of L1 as post-focus F_0 elbow

L1 in A&L 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;” (p. 55). Overall, L1 is described as exhibiting a 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” (p. 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 Figure 7, due to focus, F_0 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 already seen in Figure 7, 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 (left), but slower if it is unstressed (right). Similar stress-dependent post-focus F_0 falling speech has been found for English (Xu & Xu 2005). In other words, the ‘stress-seeking’ behaviour observed in A&L as well as other AM-based studies (Grice et al. 2000, Gussenhoven 2000, Pierrehumbert & Beckman 1988) can be accounted for in PENTA as being due to greater articulatory strength given to stressed syllables than unstressed syllables even when they are both post-focus.

4.2.3 Alignment of L2 as F_0 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” (p. 55). A&L 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.” (p. 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.” (p. 62)

To us, these patterns are again likely related to target strength due to lexical stress. That is, target strength of sentence-final syllable is dependent on lexical stress, being higher in stressed syllables and lower in unstressed syllables. The impact of this difference can be again seen in Figure 7. Both sentences have a sentence-final rising target associated with the question modality. The sentence on the left, due to the weak strength in its unstressed final syllable, shows a continuous shallow final rise. The sentence on the right, in contrast, due to the strong strength of its stressed final syllable, shows a dip in the middle of the syllable before the final rise. This dip, which is also seen in Figure 1a of A&L for the sentence ['pu 'zi] with sentence-final stress, is likely to have led to the difference in the visually marked L2 alignment in A&L. But the simulation in Figure 7 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 as reported by A&L can be accounted for by PENTA in terms of the interaction of lexical stress with focus and question intonation. With regard to scaling, A&L 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 A&L is in terms of only the F_0 peak on the wh-word and elbow of post-focus F_0 drop and sentence-final F_0 rise. From the perspective of target approximation, this lack of variability is not really surprising. As can be seen in the simulations in Figure 7, 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 Figure 7.

A&L also report that sentences that start with a stressed syllable have higher initial F_0 than those starting with an unstressed syllable. They attribute this to a truncation mechanism, with which a stressed syllable truncates a virtual L target that occurs at the left edge of every sentence in Greek. From the simulated F_0 contours in Figure 7, however, it is difficult to see how this truncation mechanism can work. Given that 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 AM theory, thus keeping the lowest initial F_0 unchanged. With the target-interpolation model, variation of initial F_0 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.

From the PENTA perspective, the idea of an utterance-initial virtual pitch is actually rather plausible, because there is already empirical evidence for it in our own data on tones produced in isolation (Xu 1997), as shown in Figure 8. We can see that different tones have different onset F_0 although the initial consonant of the syllable is [m], a sonorant. 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 F_0 from the left rather than from the right as suggested by A&L, and it would be applied regardless of whether the initial syllable is stressed.

Furthermore, because in A&L, the wh-word with initial stress starts with a voiceless consonant, ['pu], while the wh-word without initial stress starts with a vowel, [apo'pu], an F_0 contour with a rising onset is likely to start higher in the former case, as shown in Figure 9. That is, a voiceless consonant is known to perturb the F_0 contour of a syllable in two ways, raising the onset F_0 very briefly, and "truncating" an otherwise continuous F_0 movement, as can be clearly seen when compared to the F_0 of a sonorant onset (Xu & Xu 2003). Such a "truncation" mechanism is already implemented in the PENTAtrainers and tested in Xu & Prom-on (2014).

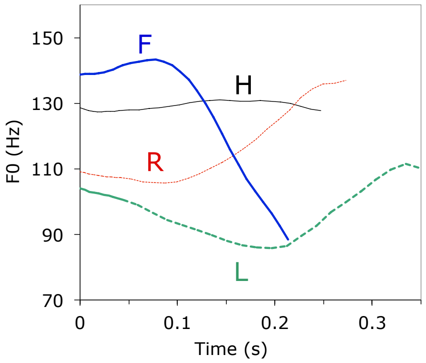


Figure 8. Mean F_0 contours of Mandarin tones in the syllable [ma] spoken in isolation by 8 speakers (averaged over 7 repetitions by 8 speakers). Data from Xu (1997).

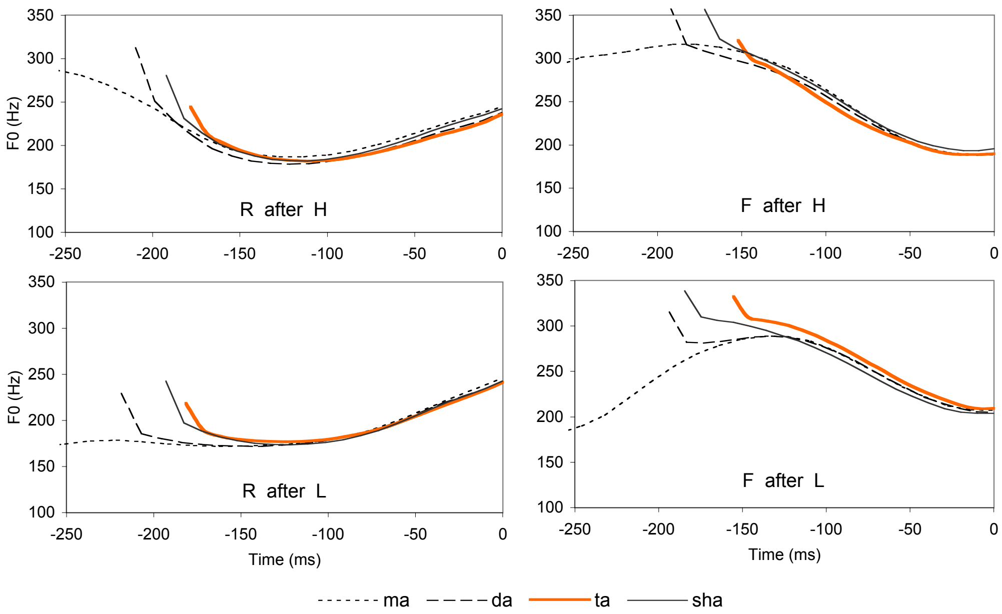


Figure 9. Effects of voiceless consonants on the F_0 contours of Mandarin R and F produced after H and L. Each curve is an average across 5 repetitions, 2 carrier sentences and 7 female speakers. All curves are aligned to the syllable offset. Data from Xu & Xu (2003).

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 are simultaneously realised through an articulatory encoding process, so that all the details of the surface prosody can be traced back to their respective sources. As such, PENTA has addressed the major criteria advocated by A&L for a complete theory of intonation, namely, *abstraction*, *generalisation*, *prediction* and *account for details*.

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 F_0 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, thus are behind language-specific variaitons; 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 et al. 2013).

Account 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. So far, a substantial amount of details in surface prosody have 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 detailed quantitative comparisons.

Acknowledgement

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

REFERENCES

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

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 & Janet Pierrehumbert (1986). Intonational structure in Japanese & English. *Phonology Yearbook* **3.** 255–309.

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

Black, Alan & Andrew Hunt (1996). Generating F0 contours from ToBI labels using linear regression. *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP 1996)*, Philadelphia, PA. 1385–1388.

Bolinger, Dwight LeMerton (1986). Intonation and its parts: Melody in spoken English. Palo Alto, CA: Stanford University Press.

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

Chen, Yiya & Yi Xu (2006). Production of weak elements in speech: Evidence from F0 patterns of neutral tone in Standard Chinese. *Phonetica* **63.** 47–75.

Doupe, Allison & Patricia 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 MacNeilage (Ed.), *Producing 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 Michael Broe & Janet Pierrehumbert (eds.), *Papers in Laboratory Phonology V: Acquisition and the lexicon.* Cambridge: Cambridge University Press. 132–151.

Gussenhoven, Carlos (2004). The phonology of tone & intonation. New York: Cambridge University Press.

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

Jun, Sun-Ah (Ed.) (2005). Prosodic typology: The phonology of intonation & phrasing. New York: Oxford University Press.

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

Ladd, D. Robert (2008). Intonational phonology. New York: Cambridge University Press.

Lee, Albert, Yi Xu & Santitham Prom-on (2014). Modeling Japanese F_0 contours using the PENTAtrainers and AMtrainer. *Proceedings of TAL2014,* Nijmegen, 164-167.

Liu, Fang & Yi Xu (2005). Parallel encoding of focus & 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 (2014). Models of vocal learning in the songbird: Historical frameworks and the stabilizing critic. *Developmental Neurobiology*, DOI:10.1002/dneu.22189.

O'Connor, Joseph & Gordon Arnold (1973). Intonation of colloquial English: A practical handbook. London: Longman.

Peng, Shu-Hui (2000). Lexical versus 'phonological' representations of Mandarin sandhi tones. In Michael Broe & Janet Pierrehumbert (eds.) *Papers in Laboratory Phonology V: Acquisition and the Lexicon*. Cambridge: Cambridge University Press, 152-167.

Pierrehumbert, Janet (1980). *The phonology & phonetics of English intonation.* PhD Thesis. Massachusetts Institute of Technology.

Pierrehumbert, Janet (1981). Synthesizing intonation. *JASA* **70**. 985–995.

Pierrehumbert, Janet (2000). Tonal elements and their alignment. In Merle Horne (ed.) *Prosody: Theory and Experiment — Studies Presented to Gösta Bruce*. London: Kluwer Academic Publishers. 11-36.

Pierrehumbert, Janet & Mary Beckman (1988). Japanese Tone Structure. Massachusetts Institute of Technology.

Pierrehumbert, Janet & Julia Hirschberg (1990). The meaning of intonational contours in the interpretation of discourse. In Philip Cohen, Jerry Morgan & Martha Pollack (Eds.), *Intentions in communication*. MIT Press. 271–311.

Prom-on, Santitham & Yi Xu (2012). PENTATrainer2: A hypothesis-driven prosody modeling tool. *Proceedings of the 5th IESL Conference on Experimental Linguistics,* Athens, Greece. 93–100.

Prom-on, Santitham, Yi Xu & Bundit Thipakorn (2009). Modeling tone & intonation in Mandarin and English as a process of target approximation. *JASA* **125.** 405–424.

Raidt, Stephan, Gérard Bailly, Bleicke Holm & Hansjörg Mixdorff (2004). Automatic generation of prosody: Comparing two superpositional systems. *Proceedings of the 2nd International Conference on Speech Prosody (SP2004)*, Nara, Japan.

Saltzman, Elliot & Kevin 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 Thesis*. 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 & Lishi Luo (1999). A pitch target approximation model for F0 contours in Mandarin. *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS 1999)*, San Francisco, CA. 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. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)*, Hong Kong. 152–155.

Xu, Yi, Szu-Wei Chen & Bei Wang (2012). Prosodic focus with & without post-focus compression: A typological divide within the same language family? *Linguistic Review* **29**. 131–147.

Xu, Yi, Andrew Kelly & Cameron Smillie (2013). Emotional expressions as communicative signals. In S. Hancil & Daniel Hirst (eds.) Prosody and Iconicity. Philadelphia: John Benjamins Publishing Co., 33-60.

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.

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 & segmental aspects of speech, In Oliver Niebuhr (Ed.), *Understanding prosody: The role of context, function and communication.* Walter de Gruyter. 1–26.

Xu, Yi & Santitham Prom-on (2010-2014). PENTAtrainer1.praat. Available from: 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 & Qi Emily Wang (2001). Pitch targets & 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.