Chapter 2
Degrees of Freedom in Prosody Modeling

Yi Xu and Santitham Prom-on

Abstract Degrees of freedom (DOF) refer to the number of free parameters in a model that need to be independently controlled to generate the intended output. In this chapter, we discuss how DOF is a critical issue not only for computational modeling, but also for theoretical understanding of prosody. The relevance of DOF is examined from the perspective of the motor control of articulatory movements, the acquisition of speech production skills, and the communicative functions conveyed by prosody. In particular, we explore the issue of DOF in the temporal aspect of speech and show that, due to certain fundamental constraints in the execution of motor movements, there is likely minimal DOF in the relative timing of prosodic and segmental events at the level of articulatory control.

2.1 Introduction

The ability to model speech prosody with high accuracy has long been the dream of prosody research, both for practical applications such as speech synthesis and recognition and for theoretical understanding of prosody. A key issue in prosody modeling is degrees of freedom (henceforth interchangeable with DOF). DOF refers to the number of independent parameters in a model that needs to be estimated in order to generate the intended output. So far there has been little serious discussion of the issue of DOF in prosody modeling, especially in terms of its theoretical implications. Nevertheless, DOF is often implicitly considered, and it is generally believed that, other things being equal, the fewer degrees of freedom in a model the better. For example, in the framework of intonational phonology, also known as the AM theory or the Pierrehumbert model of intonation, it is assumed that, at least for nontonal languages like English, "sparse tonal specification is the key to combining accurate phonetic modeling with the expression of linguistic equivalence

Y. Xu (✉)
University College London, London, UK
e-mail: yi.xu@ucl.ac.uk

S. Prom-on
King Mongkut's University of Technology, Thonburi, Thailand

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_2

of intonation contours of markedly different lengths" (Arvaniti and Ladd 2009, p. 48). The implication of such sparse tonal representation is that there is no need to directly associate F_0 events with individual syllables or words, and for specifying F_0 contours between the sparsely distributed tones. This would mean high economy of representation. Sparse F_0 specifications are also assumed in various computational models (e.g., Fujisaki 1983; Taylor 2000; Hirst 2005).

A main feature of the sparse tonal specification is that prosodic representations are assigned directly to surface F_0 events such as peaks, valleys, elbows, and plateaus. As a result, each temporal location is assigned a single prosodic representation, and an entire utterance is assigned a single string of representations. This seems to be representationally highly economical, but it also means that factors that do not directly contribute to the major F_0 events may be left out, thus potentially missing certain critical degrees of freedom. Another consequence of sparse tonal specification is that, because major F_0 events do not need to be directly affiliated with specific syllables or even words, their timing relative to the segmental events has to be specified in modeling, and this means that temporal alignment constitutes one or more (depending on whether a single point or both onset and offset of the F_0 event need to be specified) degrees of freedom. Thus many trade-offs need to be considered when it comes to determining DOF in modeling.

In this chapter we take a systematic, though brief look at DOF in prosody modeling. We will demonstrate that DOF is not only a matter of how surface prosodic events can be economically represented. Rather, decisions on DOF of a model should be ecologically valid, i.e., with an eye on what human speakers do. We advocate for the position that every degree of freedom (DOF) needs to be independently justified rather than based only on adequacy of curve fitting. In the following discussion we will examine DOF from three critical aspects of speech: motor control of articulatory movements, acquisition of speech production skills, and communicative functions conveyed by prosody.

2.2 The Articulatory Perspective

Prosody, just like segments, is articulatorily generated, and the articulatory process imposes various constraints on the production of prosodic patterns. These constraints inevitably introduce complexity into surface prosody. As a result, if not properly understood, the surface prosodic patterns due to articulatory constraints may either unnecessarily increase the modeling DOF or hide important DOF. Take F_0 for example. We know that both local contours and global shapes of intonation are carried by voiced consonants and vowels. Because F_0 is frequently in movement, either up or down, the F_0 trajectory within a segment is often rising or falling, or with an even more complex shape. A critical question from an articulatory perspective is, how does a voiced segment get its F_0 trajectory with all the fine details? One possibility is that all F_0 contours are generated separately from the segmental string of speech, as assumed in many models and theories, either explicitly or implicitly, and especially

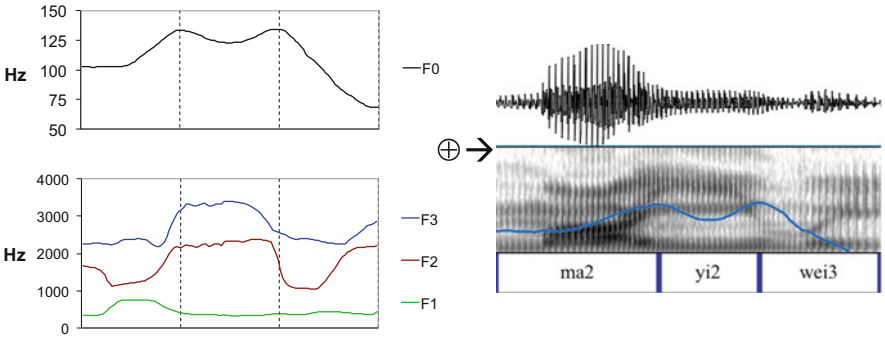

Fig. 2.1 *Left*: Continuous F_0 (*top*) and formant (*bottom*) tracks of the Mandarin utterance "(bi3) *ma2 yi2 wei3* (shan4)" [More hypocritical than Aunt Ma]. *Right*: Waveform, spectrogram and F_0 track of the same utterance. Raw data from Xu (2007)

in those that assume sparse tonal specifications (Pierrehumbert 1980; Taylor 2000; 't Hart et al. 1990). This scenario is illustrated in Fig. 2.1, where continuous F_0 and formant contours of a trisyllabic sequence are first separately generated with all the trajectory details (1a), and then merged together to form the final acoustic output consisting of both formant and F_0 trajectories. The critical yet rarely asked question is, is such an *articulate-and-merge* process biomechanically possible?

As is found in a number of studies, the change of F_0 takes a significant amount of time even if the speaker has used maximum speed of pitch change (Sundberg 1979; Xu and Sun 2002). As found in Xu and Sun (2002), it takes an average speaker around 100 ms to make an F_0 movement of even the smallest magnitude. In Fig. 2.1, for example, the seemingly slow F_0 fall in the first half of syllable 2 is largely due to a necessary transition from the high offset F_0 due to the preceding rising tone to the required low F_0 onset of the current rising tone, and such movements are likely executed at maximum speed of pitch change (Kuo et al. 2007; Xu and Sun 2002). In other words, these transitional movements are mainly due to articulatory inertia. Likewise, there is also evidence that many of the formant transitions in the bottom left panel of Fig. 2.1 are also due to articulatory inertia (Cheng and Xu 2013).

Given that F_0 and formant transitions are mostly due to inertia, and are therefore by-products of a biomechanical system, if the control signals (from the central nervous system (CNS)) sent to this system also contained all the inertia-based transitions, as shown on the left of Fig. 2.1, *the effect of inertia would be applied twice*. This consideration makes the *articulate-and-merge* account of speech production highly improbable. That is, it is unlikely that continuous surface F_0 contours are generated (with either a dense or sparse tonal specification) independently of the segmental events, and are then added to the segmental string during articulation.

But how, then, can F_0 contours and segmental strings be articulated together? One hypothesis, as proposed by Xu and Liu (2006), is that they are coproduced under the coordination of the syllable. That is, at the control level, each syllable is

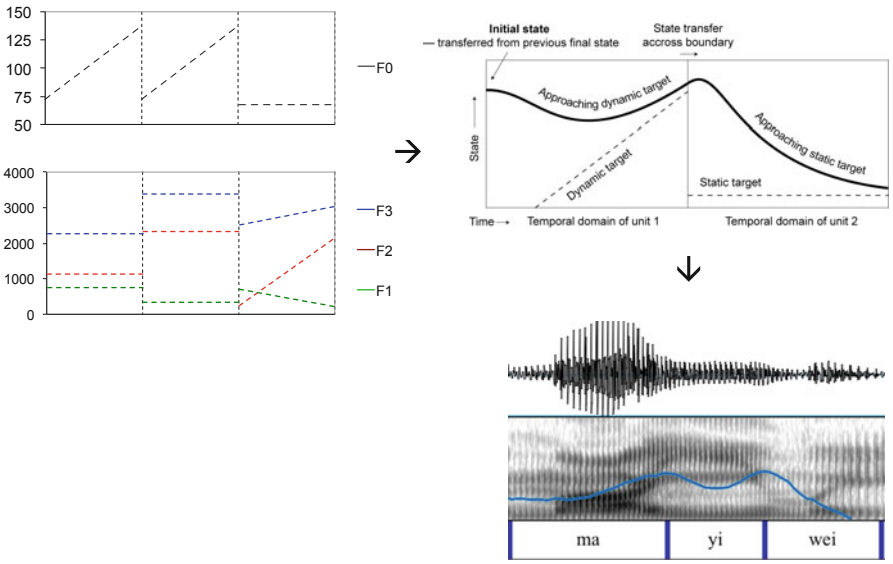

Fig. 2.2 *Left*: Hypothetical underlying pitch (*upper*) and formant (*lower*) targets for the Mandarin utterance shown in Fig. 2.1. *Top right*: The target approximation (TA) model (Xu and Wang 2001). *Bottom right*: Waveform, spectrogram and F_0 track of the same utterance. Raw data from Xu (2007)

specified with all the *underlying articulatory targets* associated with it, including segmental targets, pitch targets, and even phonation (i.e., voice quality) targets. This is illustrated in the top left block of Fig. 2.2 for pitch and formants. Here the formant patterns are representations of the corresponding vocal tract shapes, which are presumably the actual targets. The articulation process then concurrently approaches all the targets, respectively, through target approximation (top right). The target approximation process ultimately generates the continuous surface F_0 and formant trajectories (bottom right), which consist of mostly transitions toward the respective targets. Thus, every syllable, before its articulation, would have been assigned both segmental and suprasegmental targets as control signals for the articulatory system. And importantly, the effects of inertia are applied only once, during the final stage of articulatory execution.

Pitch target specification for each and every syllable may mean greater DOF than the sparse tonal specification models, of course, which is probably one of the reasons why it is not widely adopted in prosody modeling. But what may not have been apparent is that it actually reduces a particular type of DOF, namely, the F_0-segment alignment. For the sparse tonal specification models, because F_0 events are not attached to segments or syllables, the relative alignment of the two becomes a free variable, which constitutes at least one DOF (two if onset and offset of an F_0 event both have to be specified, as in the Fujisaki model). Thus for each tonal event, not only its height, but also its position relative to a segmental event, need to be specified. This complexity is further increased by the assumption of most of the sparse-tonal

specification models that the number of tonal and phrasal units is also a free variable and has to be learned or specified. For the Fujisaki model, for example, either human judgments have to be made based on visual inspection (Fujisaki et al. 2005), or filters of different frequencies are applied first to separately determine the number of phrase and accent commands, respectively (Mixdorff 2000). Even for cases where pitch specifications are obligatory for each syllable, e.g., in a tone language, there is a further question of whether there is freedom of micro adjustments of F_0-segment alignment. Allowance for micro alignment adjustments is assumed in a number of tonal models (Gao 2009; Gu et al. 2007; Shih 1986).

There has been accumulating evidence against free temporal alignment, however. The first line of evidence is the lack of micro alignment adjustment in the production of lexical tones. That is, the unidirectional F_0 movement toward each canonical tonal form starts at the syllable onset and ends at syllable offset (Xu 1999). Also the F_0-syllable alignment is not affected by whether the syllable has a coda consonant (Xu 1998) or whether the syllable–initial consonant is voiced or voiceless (Xu and Xu 2003). Furthermore, the F_0-syllable alignment does not change under time pressure, even if tonal undershoot occurs as a result (Xu 2004). The second line of evidence is from motor control research. A strong tendency has been found for related motor movements to be synchronized with each other, especially when the execution is at a high speed. This is observed in studies of finger tapping, finger oscillation, or even leg swinging by two people monitoring each other's movements (Kelso et al. 1981; Kelso 1984; Kelso et al. 1979; Mechsner et al. 2001; Schmidt et al. 1990). Even non-cyclic simple reaching movements conducted together are found to be fully synchronized with each other (Kelso et al. 1979).

The synchrony constraints could be further related to a general problem in motor control. That is, the high dimensionality of the human motor system (which is in fact true of animal motor systems in general) makes the control of any motor action extremely challenging, and this has been considered as one of the central problems in the motor control literature (Bernstein 1967; Latash 2012). An influential hypothesis is that the CNS is capable of functionally freezing degrees of freedom to simplify the task of motor control as well as motor learning (Bernstein 1967). The freezing of DOF is analogous to allowing the wheels of a car to rotate only around certain shared axes, under the control of a single steering wheel. Thus the movements of the wheels are fully synchronized, and their degrees of freedom merged. Note that such synchronization also freezes the relative timing of the related movements, hence eliminating it as a DOF. This suggests that the strong synchrony tendency found in many studies (Kelso et al. 1979; Kelso et al. 1981; Mechsner et al. 2001; Schmidt et al. 1990) could have been due to the huge benefits brought by the reduction of temporal degrees of freedom.

The benefit of reducing temporal DOF could also account for the tone-syllable synchrony in speech found in the studies discussed above. Since articulatory approximations of tonal and segmental targets are separate movements that need to be produced together, they are likely to be forced to synchronize with each other, just like in the cases of concurrent nonspeech motor movements. In fact, it is possible that

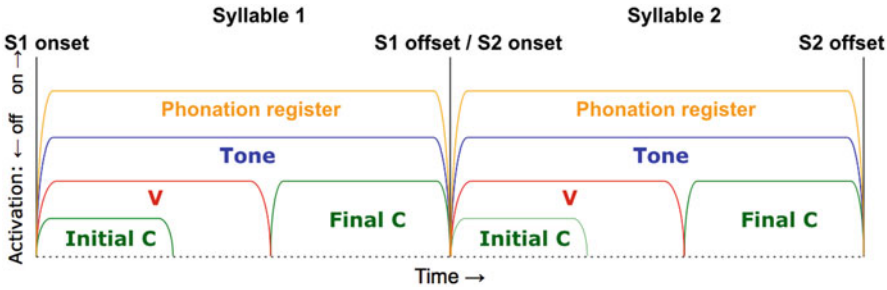

Fig. 2.3 The time structure model of the syllable (Xu and Liu 2006). The syllable is a *time structure* that assigns temporal intervals to consonants, vowels, tones, and phonation registers (each constituting a phone). The alignment of the temporal intervals follows three principles: **a** *Co-onset* of the initial consonant, the first vowel, the tone, and the phonation register at the beginning of the syllable; **b** *Sequential offset* of all noninitial segments, especially coda C; and **c** *Synchrony* of laryngeal units (tone and phonation register) with the entire syllable. In each case, the temporal interval of a phone is defined as the time period during which its target is approached

the syllable is a mechanism that has evolved to achieve synchrony of multiple articulatory activities, including segmental, tonal, and phonational target approximations. As hypothesized by the *time structure model of the syllable* (Xu and Liu 2006), the syllable is a temporal structure that controls the timing of all its components, including consonant, vowel, tone, and phonation register (Xu and Liu 2006), as shown in Fig. 2.3. The model posits that the production of each of these components is to articulatorily approach its ideal target, and the beginning of the syllable is the onset of the target approximation movements of most of the syllabic components, including the initial consonant, the first vowel, the lexical tone, and the phonation register (for languages that use it lexically). Likewise, the end of the syllable is the offset of all the remaining movements. In this model, therefore, there is always full synchrony at the onset and offset of the syllable. Within the syllable, there may be free timing at two places, the offset of the initial consonant, and the boundary between the nuclear vowel and the coda consonant. In the case of lexical tone, it is also possible to have two tonal targets within one syllable, as in the case of the L tone in Mandarin, which may consist of two consecutive targets when said in isolation. The boundary between the two targets is probably partially free, as it does not affect synchrony at the syllable edges.

2.3 The Learning Perspective

Despite the difficulty of motor control just discussed, a human child is able to acquire the ability to speak in the first few years of life, without formal instructions, and without direct observation of the articulators of skilled speakers other than the visible ones such as the jaw and lips. The only sure input that can inform the child of the articulatory details is the acoustics of speech utterances. How, then, can the child

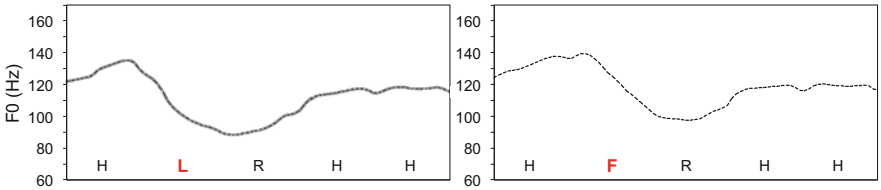

Fig. 2.4 Mean time-normalized F_0 contours of five-syllable Mandarin sentences with syllable two carrying the low tone on the *left* and the falling tone on the *right*. Data from Xu (1999)

learn to control her own articulators to produce speech in largely the same way as the model speakers? One possibility is that the acquisition is done through analysis-by-synthesis with acoustics as the input. The strategy is also known as distal learning (Jordon and Rumelhart 1992). To be able to do it, however, the child has to face the problem of multiplicity of DOF. As discussed earlier, adult speech contains extensive regions of transitions due to inertia. Given this fact, how can the child know which parts of the surface contour are mostly transitions, and which parts best reflect the targets? In Fig. 2.4, for example, how can a child tell that the Mandarin utterance on the left contains a low tone, while the one on the right contains a falling tone at roughly the same location? One solution for the child is to confine the exploration of each tonal target to the temporal domain of the syllable. That way, the task of finding the underlying target is relatively simple. This strategy is implemented in our computational modeling of tone as well as intonation (Liu et al. 2013; Prom-on et al. 2009; Xu and Prom-on 2014). Our general finding is that, when the syllable is used as the tone-learning domain, their underlying targets are easily and accurately extracted computationally, judging from the quality of synthesis with the extracted tonal targets in all these studies.

The ease of extracting tonal targets within the confine of the syllable, however, does not necessarily mean that it is the best strategy. In particular, what if the synchronization assumption is relaxed so that the learning process is given some freedom in finding the optimal target-syllable alignment? In the following we will report the results of a modeling experiment on the effect of flexibility of timing in pitch target learning.

2.3.1 Effect of Freedom of Tone–Syllable Alignment on Target Extraction—An Experiment

The goal of this experiment is to test if relaxing strict target-syllable synchrony improves or reduces F_0 modeling accuracy and efficiency with an articulatory-based model. If there is real timing freedom either in production or in learning, modeling

Fig. 2.5 Illustration of *onset timing shifts* used in the experiment and their impacts on the timing of adjacent syllables

accuracy should improve with increased timing flexibility during training. Also assuming that there is regularity in the target alignment in mature adults' production, the process should be able to learn the alignment pattern if given the opportunity.

2.3.1.1 Method

To allow for flexibility in target alignment, a revised version of PENTAtrainer1 (Xu and Prom-on 2010–2014) was written. The amount of timing freedom allowed was limited, however, as shown in Fig. 2.5. Only onset alignment relative to the original was made flexible. For each syllable, the onset of a pitch target is either set to be always at the syllable onset (fixed alignment), or given a 50 or 100 ms search range (flexible alignment). In the case of flexible alignments, if the hypothetic onset is earlier than the syllable onset, as shown in row two, the synthetic target approximation domain becomes longer than that of the syllable, and the preceding target domain is shortened; if the hypothetic onset is later than the syllable onset, as shown in row three, the synthetic target approximation domain is shortened, and the preceding domain is lengthened. Other, more complex adjustment patterns would also be possible, of course, but even from this simple design, we can already see that the adjustment of any particular alignment has various implications not only for the current target domain, but also for adjacent ones.

The training data are from Xu (1999), which have been used in Prom-on et al. (2009). The dataset consists of 3840 five-syllable utterances recorded by four male and four female Mandarin speakers. In each utterance, the first two and last two syllables are disyllabic words while the third syllable is a monosyllabic word. The first and last syllables in each sentence always have the H tone while the tones of the other syllables vary depending on the position: H, R, L, or F in the second syllable, H, R, or F in the third syllable, and H or L in the fourth syllable. In addition to tonal variations, each sentence has four focus conditions: no focus, initial focus, medial focus, and final focus. Thus, there are 96 total variations in tone and focus.

2.3.1.2 Results

Figure 2.6 displays bar graphs of RMSE and correlation values of resynthesis performed by the modified PENTAtrainer1 using the three onset time ranges. As can be seen, the 0 ms condition produced lower RMSE and higher correlation than

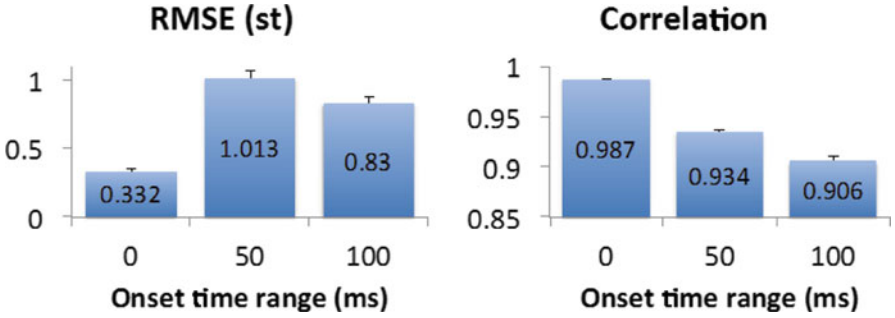

Fig. 2.6 Root mean square error (*RMSE*) and Pearson's correlation in resynthesis of F_0 contours of Mandarin tones in connected speech (data from Xu 1999) using targets obtained with three onset time ranges: 0, 50, and 100 ms

both the 50 and 100 ms conditions. Two-way repeated measures ANOVAs showed a highly significant effect of onset time range on both RMSE ($F\ [2,\ 7] = 387.4$, $p < 0.0001$) and correlation ($F\ [2,\ 7] = 320.1$, $p < 0.0001$). Bonferroni/Dunn post-hoc tests showed significant differences between all onset time ranges for both RMSE and correlation. More interestingly, on average, the learned alignments in the flexible conditions are still centered around syllable onset. The average deviation from the actual syllable boundaries is -2.3 ms in the 50 ms onset range condition and -5.1 ms in the 100 ms onset range condition (where the negative values mean that the optimized onset is earlier than the syllable boundary). A similarly close alignment to the early part of the syllable has also been found in Cantonese for the accent commands in the Fujisaki model, despite lack of modeling restrictions on the command–syllable alignment (Gu et al. 2007).

Figure 2.7 shows an example of curve fitting with 0 and 100 ms onset shift ranges. As can be seen, pitch targets learned with the 0 onset shift range produced a much tighter curve fitting than those learned with free timing. More importantly, we can see why the increased onset time range created problems. For the third syllable, the learned optimal onset alignment is later than the syllable onset. As a result, the temporal interval for realizing the preceding target is increased, given the alignment adjustment scheme shown in Fig. 2.6. As a result, the original optimal F_0 offset is no longer optimal, which leads to the sizeable discrepancy in Fig. 2.7b. Note that it is possible for us to modify the learning algorithm so that once an optimized onset alignment deviates from the syllable boundary, the preceding target is reoptimized. However, that would lead to a number of other issues. Should this reoptimization use fixed or flexible target onset? If the latter, shouldn't the further preceding target also be reoptimized? If so, the cycles will never end. Note also, that having a flexible search range at each target onset already increases the number of searches by many folds; having to reapply such searches to earlier targets would mean many more folds of increase. Most importantly, these issues are highly critical not just for modeling, but also for human learners, because they, too, have to find the optimal targets during their vocal learning.

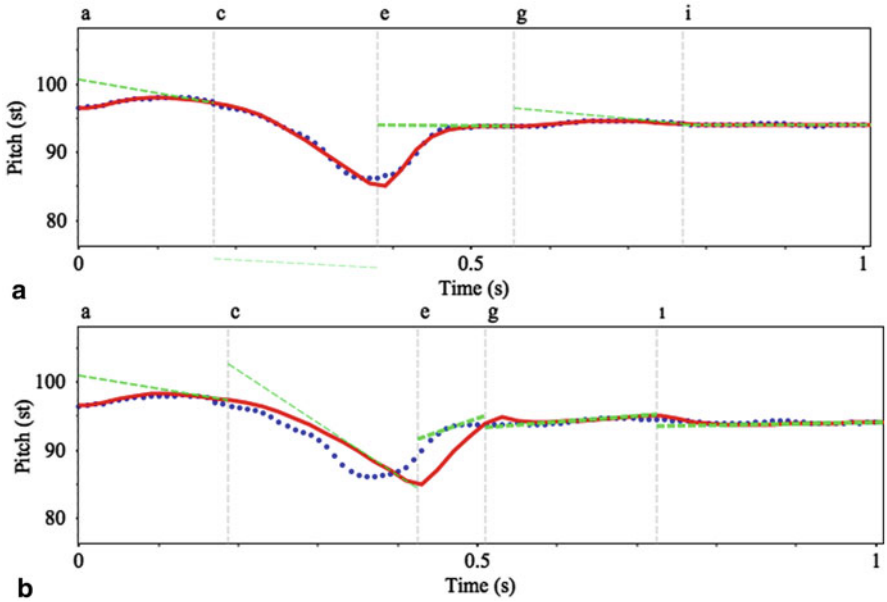

Fig. 2.7 Examples of curve fitting with targets learned with 0 ms onset timing shift (**a**), and 100 ms shift (**b**). The *blue dotted lines* are the original contours and the *red solid lines* the synthetic ones

In summary, the results of this simple modeling experiment demonstrate the benefit of fixing the temporal domain of the tonal target to that of the syllable in tone learning. Fully synchronizing the two temporal domains reduces DOF, simplifies the learning task, shortens the learning time, and also produces better learning results.

2.4 The Functional Perspective

Given that prosody is a means to convey communicative meanings (Bailly and Holm 2005; Bolinger 1989; Hirst 2005), the free parameters in a prosody model should be determined not only by knowledge of articulatory mechanisms, but also by consideration of the communicative functions that need to be encoded. Empirical research so far has established many functions that are conveyed by prosody, including lexical contrast, focus, sentence type (statement versus question), turn taking, boundary marking, etc. (Xu 2011). Each of these functions, therefore, needs to be encoded by specific parameters, and all these parameters would constitute separate degrees of freedom. In this respect, a long-standing debate over whether prosody models should be linear or superpositional is highly relevant. The linear approach, as represented by the autosegmental–metrical (AM) theory (Ladd 2008; Pierrehumbert 1980; Pierrehumbert and Beckman 1988), is based on the observation that speech intonations manifest clearly visible F_0 peaks, valleys, and plateaus. It is therefore assumed that

prosody consists of strings of discrete prosodic units, each exclusively occupying a temporal location. Such a linear approach naturally requires rather limited degrees of freedom.

The superpositional models, on the other hand, assume that surface F_0 contours are decomposable into *layers*, each consisting of a string of F_0 shapes, and the shapes of all the layers are added together to form surface F_0 contours (Bailly and Holms 2005; Fujisaki 1983; Thorsen 1980; van Santen et al. 2005). Take the Fujisaki model, for example, two layers are used to represent local shapes generated by accent commands, and global shapes generated by phrase commands, respectively. The output of the two layers is added together on a logarithmic scale to form a smooth global surface F_0 contour. Thus, superpositional models allow more than one unit to occur at any particular temporal location. This means more DOF than the linear models. In terms of economy of DOF, therefore, superpositional models may seem less optimal than linear models.

However, economy of DOF should not be the ultimate goal of prosody modeling. Instead, a model should be able to represent as many meanings conveyed by prosody as possible, while minimizing redundancy of representation. From this perspective, superpositional models with more than one layer of potential prosodic unit are aiming to provide sufficient DOF for encoding rich prosodic meanings (e.g., Bailly and Holm 2005), which makes them compare favorably to linear models. Meanwhile, however, as shown in our earlier discussion on articulatory mechanisms, each and every DOF should be articulatorily plausible. In this regard, superposition models still share with linear models the problematic *articulate-and-merge* assumption. Furthermore, from a modeling perspective, a superposition model has to first separate the surface contours into different layers, each corresponding to a particular communicative function. But this task is by no means easy. In Mixdorff (2000), filters of different frequencies were used to first separate surface F_0 contours into large global waves and small local ripples. Phrase commands are then sought from the global waves and accent commands from the local ripples. But the results are not satisfactory and manual intervention is often still needed (Mixdorff 2012).

The parallel encoding and target approximation model (PENTA) takes both articulatory mechanisms and communicative functions into consideration (Xu 2005). As shown in Fig. 2.8, the articulatory mechanism that generates surface F_0 is syllable-synchronized sequential target approximation, shown in the lower panel of the figure. In this mechanism, each syllable is assigned an underlying pitch target, and surface F_0 is the result of continuous articulatory approximation of successive pitch targets. This process is controlled by four free parameters: target height, target slope, target approximation rate, and duration. Each of these parameters therefore constitutes a DOF controllable by the encoding schemes. But there is no temporal DOF, as each target approximation is fully synchronized with the associated syllable. The encoding schemes each correspond to a specific communicative function, and the communicative functions are assumed to be parallel to (rather than dominating) each other, hence "parallel" in the name of the model. Like superposition, parallel encoding allows more than one prosodic element at any temporal location. But unlike superposition, in which streams of surface output are generated separately and then

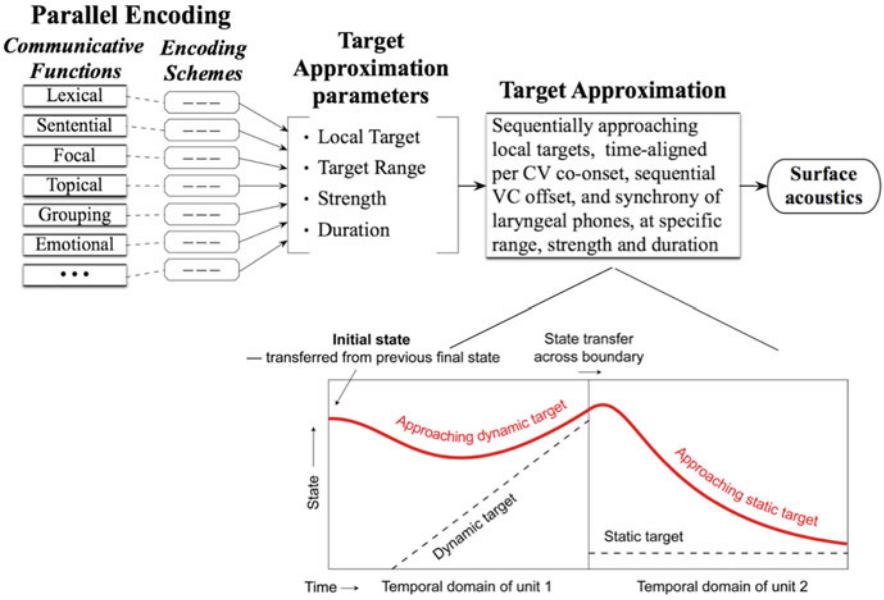

Fig. 2.8 *Upper panel*: A schematic sketch of the PENTA model (Xu 2005). *Lower panel*: The target approximation model of the articulation process (Xu and Wang 2001)

combined by summation, the encoding schemes in PENTA all influence a common sequence of underlying targets. The final sequence of targets carrying the combined influences from all communicative functions then generate surface output in a single articulatory process.

This single-target-sequence assumption of PENTA not only makes the generation of surface F_0 contours a rather straightforward process, it also makes it easy to account for a particular type of prosodic phenomenon, namely, target shift under the interaction of different communicative functions. Target shift is most vividly seen in the case of tone sandhi, whereby the underlying pitch target of a lexical tone is changed from the canonical one to a form that is very different in certain contextual conditions. In Mandarin, for example, the Low tone changes its target from low-level to rising when followed by another Low tone, and the resulting surface F_0 closely resembles that of the Rising tone. In American English, the pitch target of a word–final stressed syllable is a simple high in a pre-focus position; it changes to a steep fall when under focus in a statement, but to a steep rise in a question (Liu et al. 2013; Xu and Prom-on 2014; Xu and Xu 2005). In the PENTA approach, such a shift is modeled without taking any special steps, since for each functional combination a unique target has to be learned whether or not a sizable target shift occurs. Therefore, to the extent the PENTA approach is ecologically realistic, the way it models target shift may suggest why target shifts occur in languages in the first place. That is, because each multifunctional target needs to be learned as a whole, it is possible for

them to evolve properties that deviate significantly from their canonical forms. This issue is worth further exploration in future studies.

In terms of the specific target parameters in a model, the justification for each should come from empirical evidence of their usage in a specific function, and this principle is followed in the development of the PENTA model. For example, lambda, the rate of target approximation, could be fixed just like the time constant in the Fujisaki model (Fujisaki 1983). However, empirical research has provided evidence that the neutral tone in Mandarin and unstressed syllable in English approach their targets at much slower rates than normal tones and stressed syllables (Chen and Xu 2006; Xu and Xu 2005). Modeling studies have also shown that much lower lambda values are learned for the neutral tone and unstressed syllables (Prom-on et al. 2012; Xu and Prom-on 2014). Thus there is sufficient justification to keep lambda as a free parameter. Likewise, there is both analytical and modeling evidence for Rising and Falling tones to have unitary dynamic targets rather than successive static targets (Xu 1998; Xu and Wang 2001). Target duration is found to be used mainly in boundary marking, lexical contrast, and focus (Xu and Wang 2009; Wagner 2005). Target slope is found to be critical for tonal contrast.

In the PENTA approach, therefore, although there are only four free parameters at the target level, at the functional level, there can be as many degrees of freedom as required by the number of functions and the number of function-internal categories that need to be encoded. For English, for example, the communicative functions that need to be prosodically encoded include lexical stress, focus, sentence type, boundary marking, etc. (Liu et al. 2013; Xu and Xu 2005). For focus, it is necessary to control separate target attributes in pre-focus, on-focus, and post-focus regions. For sentence type, target attributes need to be controlled throughout the sentence, especially from the focused location onward (Liu et al. 2013). Also, focus, sentence type, and lexical stress have three-way interactions that determine the final attributes of all pitch targets in a sentence, which often result in rich diversities in local pitch targets (Liu et al. 2013).

2.5 Summary

We have examined the issue of degrees of freedom in speech prosody modeling from three different perspectives: the motor control of articulatory movements, the acquisition of speech production skills, and the communicative functions conveyed by prosody. From the articulatory perspective, we have shown that it is unlikely for the CNS to first generate separate continuous laryngeal and superlaryngeal trajectories and then merge them together when producing the whole utterance. Rather, it is more likely that individual syllables are assigned underlying laryngeal and supralaryngeal targets before their execution; and during articulation, multiple target approximation movements are executed in synchrony, under the time structure provided by the syllable. From the learning perspective, a new modeling experiment demonstrated the benefit of having minimum temporal DOF when learning pitch targets from

continuous speech, i.e., by confining the target search strictly within the temporal domain of the syllable. From the functional perspective, we have demonstrated how the PENTA approach allows multiple encoding schemes of prosodic functions to influence a common string of underlying targets, and then generate surface output in a single articulatory process of syllable synchronized sequential target approximation. We have further argued that DOF at the functional level should be based on the number of functions and number of function-internal categories that need to be encoded.

Overall, we have shown that DOF is a critical issue not only for the computational modeling of prosody, but also for the theoretical understanding of how speech prosody, and probably speech in general, can be learned in acquisition and articulated in skilled production.

References

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

Bailly, G., and B. Holm. 2005. SFC: A trainable prosodic model. *Speech Communication* 46:348–364.

Bernstein, N. A. 1967. *The co-ordination and regulation of movements*. Oxford: Pergamon.

Bolinger, D. 1989. *Intonation and its uses—melody in grammar and discourse*. California: Stanford University Press.

Chen, Y., and Y. Xu. 2006. Production of weak elements in speech—evidence from f0 patterns of neutral tone in standard chinese. *Phonetica* 63:47–75.

Cheng, C., and Y. Xu 2013. Articulatory limit and extreme segmental reduction in Taiwan Mandarin. *Journal of the Acoustical Society of America* 134 (6):4481–4495.

Fujisaki, H. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In *The production of speech,* ed. P. F. MacNeilage, 39–55. New York: Springer-Verlag.

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

Gao, M. 2009. Gestural coordination among vowel, consonant and tone gestures in Mandarin Chinese. *Chinese Journal of Phonetics* 2:43–50.

Gu, W., K. Hirose, and H. Fujisaki. 2007. Analysis of tones in Cantonese speech based on the command-response model. *Phonetica* 64:29–62.

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

Jordan, M. I., and D. E. Rumelhart. 1992. Forward models: Supervised learning with a distal teacher. *Cognitive Science* 16:316–354.

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

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

Kelso, J. A. S., K. G. Holt, P. Rubin, and P. N. Kugler. 1981. Patterns of human interlimb coordination emerge from the properties of non-linear, limit cycle oscillatory processes: Theory and data. *Journal of Motor Behavior* 13:226–261.

Kuo, Y.-C., Y. Xu, and M. Yip. 2007. The phonetics and phonology of apparent cases of iterative tonal change in standard chinese. In *Tones and tunes Vol. 2: Experimental studies in word and sentence prosody*, ed. C. Gussenhoven and T. Riad, 211–237. Berlin: Mouton de Gruyter.

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

Latash, M. L. 2012. Fundamentals of motor control. London: Academic Press.
Liu, F., Y. Xu, S, Prom-on, and A. C. L. Yu. 2013. Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modeling. *Journal of Speech Sciences* 3 (1): 85–140.
Mechsner, F., D. Kerzel, G. Knoblich, and W. Prinz. 2001. Perceptual basis of bimanual coordination. *Nature* 414:69–73.
Mixdorff, H. 2000. A novel approach to the fully automatic extraction of fujisaki model parameters. In *Proceedings of ICASSP 2000*, Istanbul, Turkey, 1281–1284.
Mixdorff, H. 2012. The application of the Fujisaki model in quantitative prosody research. In *Understanding prosody—The role of context, function, and communication,* ed. O. Niebuhr, 55–74. New York: Walter de Gruyter.
Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*. PhD. Diss., MIT, Cambridge, MA. (Published in 1987 by Indiana University Linguistics Club, Bloomington).
Pierrehumbert, J., and M. Beckman. 1988. *Japanese tone structure*. Cambridge: The MIT Press.
Prom-on, S., Y. Xu, and B. Thipakorn. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America* 125:405–424.
Prom-on, S., F. Liu, and Y. Xu 2012. Post-low bouncing in Mandarin chinese: Acoustic analysis and computational modeling. *Journal of the Acoustical Society of America*. 132:421–432.
Schmidt, R. C., C. Carello, and M. T. Turvey. 1990. Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance* 16:227–247.
Shih, C. 1986. The prosodic domain of tone sandhi in Chinese. PhD. Diss., University of California, San Diego.
Sundberg, J. 1979. Maximum speed of pitch changes in singers and untrained subjects. *Journal of Phonetics* 7:71–79.
't Hart, J., R. Collier, and A. Cohen. 1990. *A perceptual study of intonation—An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.
Taylor, P. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America* 107:1697–1714.
Thorsen, N. G. 1980. A study of the perception of sentence intonation—Evidence from Danish. *Journal of the Acoustical Society of America* 67:1014–1030.
van Santen, J., A. Kain, E. Klabbers, and T. Mishra. 2005. Synthesis of prosody using multi-level unit sequences. *Speech Communication* 46:365–375.
Wagner, M. 2005. Prosody and recursion. PhD. Diss., Massachusetts Institute of Technology.
Xu, Y. 1998. Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica* 55:179–203.
Xu, Y. 1999. Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics* 27:55–105.
Xu, Y. 2004. Understanding tone from the perspective of production and perception. *Language and Linguistics* 5:757–797.
Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46:220–251.
Xu, Y. 2007. Speech as articulatory encoding of communicative functions. Proceedings of the 16th international congress of phonetic sciences, Saarbrucken, 25–30.
Xu, Y. 2011. Speech prosody: A methodological review. *Journal of Speech Sciences* 1:85–115.
Xu, Y., and F. Liu. 2006. Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics* 18:125–159.
Xu, Y., and S. Prom-on. 2010–2014. PENTAtrainer1.praat. http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1/. Accessed 24 Nov 2013.
Xu, Y., and S. 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, Y., and X. Sun. 2002. Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America* 111:1399–1413.

Xu, Y., and M. Wang. 2009. Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics* 37:502–520.

Xu, Y., and Q. E. Wang. 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* 33:319–337.

Xu, C. X., and Y. Xu 2003. Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association* 33:165–181.

Xu, Y., and C. X. Xu. 2005. Phonetic realization of focus in English declarative intonation. *Journal of Phonetics* 33:159–197.