Articulatory-Functional Modeling of Speech Prosody: A Review

Yi Xu[1], *Santitham Prom-on*[2]

[1] Department of Speech, Hearing and Phonetic Sciences, University College London, UK
[2] Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand
yi.xu@ucl.ac.uk, santitham@cpe.kmutt.ac.th

Abstract

Natural prosody is produced by an articulatory system to convey communicative meanings. It is therefore desirable for prosody modeling to represent both articulatory mechanisms and communicative functions. There are doubts, however, as to whether such representation is necessary or beneficial if the aim of modeling is to just generate perceptually acceptable output. In this paper we briefly review models that have attempted to implement representations of either or both aspects of prosody. We show that, at least theoretically, it is beneficial to represent both articulatory mechanisms and communicative functions even if the goal is to just simulate surface prosody.

Index Terms: speech prosody, modeling, PENTA, qTA

1. Introduction

In speech, prosody plays an important role in conveying both communicative meanings and individual characteristics. Modeling speech prosody is thus important not only for theoretical understanding of speech phenomena, but also for the advancement of speech technologies. For speech synthesis, prosody generation is still a major bottleneck that limits the naturalness of text-to-speech applications. For speech recognition, proper extraction of information encoded in speech prosody is very important not only for applications like speech-to-concept systems, but also for proper processing of lexical information such as tone or word stress.

Most existing approaches to prosody modeling try to generate directly observable acoustic events such as the location, height and shape of F_0 peaks and valleys, stylized piece-wise linear contours, or individual F_0 points [1-4]. The disadvantage of such direct acoustic feature modeling is that either a large amount of data points have to be stored, or arbitrary (thus unnatural) transitional functions have to be employed. The alternative is to take the process of F_0 production into consideration and try to simulate the underlying mechanisms of speech articulation, e.g., [5-6]. A potential advantage of articulatory-based modeling is that parameters are much more simplified without loss of naturalness.

Beside the need to model articulatory mechanisms, the way of coding the communicative meanings and expressions is also important. The current common practice is to mark up surface prosodic events such as pitch accents, boundary tones and break indices [3-4,7-8]. The problem is that these events may not directly correspond to specific communicative meanings [9-11]. Attempts to model communicative functions more directly have been made [1,9-10], but so far they have yet to be translated into robust improvement to speech technology.

This paper presents a review of the speech prosody models that have attempted to model either articulatory or functional aspects of prosody. Section 2 discusses the need to implement mechanisms related to prosody production and models that have tried to implement such representations. Section 3 discusses motivations for directly representating communicative functions and models that implement such representation.

2. Articulatory Mechanisms

One of the key features of a good prosody model is the ability to accurately simulate the acoustic events. To achieve this goal, there are two possible approaches; one is to directly model the surface acoustic events and the other is to model the articulatory process that generates such acoustic events. Models utilizing the concept of direct modeling of acoustic events are derived mainly based on the shape of the F_0 contours, with minimal consideration about the articulatory process of F_0 production. The examples of models using this approach are the quadratic spline model [2], the Pierrehumbert model [3], the tilt model [4], the linear alignment model [8], and the superposition of functional contours (SFC) model [1]. The quadratic spline model interpolates peaks and valleys of F_0 contours with a quadratic spline function while the Pierrehumbert model interpolates F_0 between adjacent peaks and valleys using a linear or sagging function. The tilt model generates F_0 from the tilt parameters which describe the shapes of F_0 in each intonational event, e.g., pitch accent and boundary tone. The F_0 contour of an utterance is represented by a series of these intonational events. The linear alignment model uses curve classes as templates, warping and then combining these curve classes superpositionally to generate F_0 contours. The SFC model simulates intonation by combining multiple elementary contours that are functionally defined. Although models in this category can represent F_0 contours at a certain level of accuracy, they do not separate surface patterns that carry intended information from those that are due to articulatory mechanisms. As a result, they have to either ignore most of the variations due to articulation, as done in various stylization strategies, or simulate all surface F_0 patterns directly as just described.

A number of researchers have taken a different approach to the modeling problem. Instead of controlling the surface acoustic features directly, they proposed models that focus on simulating the articulatory process. The acoustic features in this case are treated as the outcome of the model simulation. Examples in this category are the soft-template model [6,12] and the command-response model [5,13]. The soft-template markup language (Stem-ML), based on a soft-template model, describes F_0 contours as resulting from realizing underlying tonal templates with different amounts of muscle forces under the physical constraint of smoothness [12]. The smoothness constraint guarantees continuous connections between adjacent templates, and the varying muscle force determines the degree to which the shape of each template is preserved in the surface F_0 under the influence of neighboring tones that are either adjacent or far away, and either preceding or following the targeted template. Stem-ML uses the optimization modeling approach for F_0 realization which requires

sophisticated and complex error minimization. Even though the assumptions of Stem-ML are motivated by physical mechanisms, it requires complex mathematical translation from articulatory constraints to effort and error constraints in the optimization. The command-response model [5,13] represents surface F_0 as the logarithmic sum of phrase components and accent or tone components. Each of these components controls the F_0 variation in global and local scale respectively. The basic idea of the command-response model is to model the muscular commands and their responses. Thus, each of the commands corresponds to an individual muscular command.

2.1. Target Approximation

The Target Approximation (TA) model started with the observation that variable surface F_0 patterns of lexical tones seem to have consistent underlying constituents [14-15] which behave like ideal pitch patterns that speakers try to achieve during individual syllables [16]. The variability, which can at times very extensive [14-15,17-18], seems to be closely related to the limit of the maximum speed of pitch change speakers can achieve [19]. Because of such variability, observed F_0 contours cannot directly correspond to the communicative meanings but seem to be the output of a target approximation process, as illustrated in Figure 1. The theoretical TA framework has been recently implemented computationally as the quantitative target approximation (qTA) model [20], which simulates an articulatory-oriented F_0 control mechanism for generating tone and intonation. The core of qTA is the target approximation mechanism as depicted in Figure 1. In qTA, a pitch target is a forcing function representing the joint muscular force of the laryngeal muscles that control vocal fold tension. qTA has been tested numerically by using it to simulate tone, lexical stress and focus in Mandarin and English with an automatic analysis-by-synthesis procedure [20]. Comparisons of qTA generated F_0 contours with those of natural speech showed encouraging results in terms of rmse, correlation, perceptual identification of tone and focus and judgment of naturalness [20].

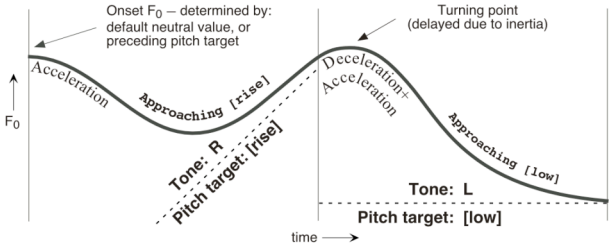

Figure 1: Illustration of the TA model. The vertical lines represent syllable boundaries. The dashed lines represent underlying pitch targets. The thick curve represents the F_0 contour that results from asymptotic approximation of the pitch targets [16].

In qTA a pitch target is represented by a simple linear equation,

$$x(t) = mt + b \tag{1}$$

where m and b denote the slope and height of the pitch target, respectively.

The control of the vocal fold tension in qTA is implemented through a third-order critically damped linear system, in which the total response is

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

where the first term, $x(t)$, is the forced response of the system which is the pitch target and the second term is the natural response. The transient coefficients c_1, c_2, and c_3 are determined jointly by the initial conditions and the target of the articulatory process. The initial conditions are the initial state of the dynamic F_0 movement, consisting of initial F_0 level, $f_0(0)$, initial velocity, $f_0'(0)$, and initial acceleration, $f_0''(0)$. Solving the systems of linear equations determined from the initial conditions, the transient coefficients can be computed with the following formulae.

$$c_1 = f_0(0) - b \tag{3}$$

$$c_2 = f_0'(0) + c_1\lambda - m \tag{4}$$

$$c_3 = \left(f_0''(0) + 2c_2\lambda - c_1\lambda^2\right)/2 \tag{5}$$

2.2. Comparison of qTA with other models

qTA has been tested only on limited amount of data in Mandarin and English and no direct comparisons have been made between its performance and those of other models. What we would like to do here is to highlight the key assumptions behind qTA and how they differ from those of other models, with due recognition that differences do not necessarily mean advantages.

From certain perspective, one could view qTA as yet another attempt to stylize the F_0 contour using certain functions. However, the conceptualization of qTA is based on an explicit set of assumptions derived from prior empirical research:

1. F_0 generation via vocal fold tension control;
2. Pitch targets as the basic control unit/level;
3. pitch targets are either static or dynamic
4. Unidirectional sequential target approximation
5. State transfer across target approximation movements;
6. Syllable synchronization.

Assumption 1 is similar to the assumption of the Fujisaki model, which recognizes that surface F_0 cannot be controlled directly, but rather through the manipulation of vocal fold tension which is directly proportional to the logarithmic scale of F_0. Assumption 2 is about the most basic level of pitch control. For this qTA differs from the Fujisaki model in that it tries to model only the joint force of all laryngeal muscles that control the tension of the vocal folds instead of modeling the actions of individual muscles. This is based on the knowledge that muscles are controlled in functional groups rather than individually [21], Note that this is also an issue of economy of modeling. Simulation at the level of individual muscular forces would entail greater degree of freedom, as the individual muscular forces have to be differentially adjusted according to the distance to be covered between the initial and targeted articulatory states.

Assumption 3 recognizes that targets themselves can be either static or dynamic. Much debate has taken place in both segmental and prosodic aspects of speech as to whether movements or contours are intrinsic to the basic speech units [22-23], but what is not generally recognized is that both static and dynamic targets could generate surface movements, but they may differ in terms of the detailed dynamics they produce. Note that, this assumption is not fully unique to qTA. The possibility of having underlyingly dynamic components is implicitly recognized in the Fujisaki model, the SFC model and the Stem-ML model.

According to assumption 4 all movements unidirectionally approach one target or another in sequence. This means that there is no return phases to a base line or a neutral position, which is assumed to either obligatory or optional in other models based on a damped linear system [5,24]. It also means that targets influence each other only from left to right, which

differs from the Stem-ML model which assumes symmetrical influence targets in both temporal directions.

Assumption 5 is unique to qTA as it is not implemented in any other model based on a damped linear system, at least to our knowledge. In both Fujisaki model and the Task Dynamic model, it is assumed that a command or a equilibrium point is either fully achieved [5,13] or reached a quasi-static point (task dynamic), and the only state transfer across movements is that of displacement, since velocity and acceleration would have been reached 0 by the end of a command or gesture. Transfer of velocity and acceleration across movements is not just for the sake of guaranteeing smoothness at the junctions, which may not even be perceptually important, but to simulate carryover influences which sometimes can be quite extensive. In fact, it has been found that by the middle of a syllable, the influence of the preceding tone is mostly in terms of the final velocity rather than final F_0 of the preceding tone [25]. Figure 2 is a qTA simulation of the Mandarin tone sequences RNNF (solid) and HNNF (dotted), where R stands for the rising tone, H the high tone, N the neutral tone and F the falling tone. As can be seen, the final F_0 of R is much lower than that of H, but the F_0 of the NN sequence after R has surpassed that of the NN sequence after H. And the remnant of the effect is visible even at the beginning of F.

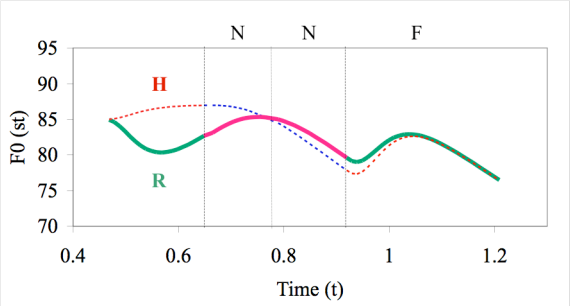

Figure 2. Simulation of Mandarin tone sequences HNNF and RNNF (spoken by author YX). The values of *m*, *b* and λ are 50, 40, -12 for R, 0, 2, 40 for H, 0, -20, 9 for N and -50, 6, 37 for F.

The final assumption, namely, syllable synchronization, is one of the most questioned ones, especially when it is applied to a non-tonal language like English. Note first that this is actually an implementational assumption in qTA, because it is not defined as part of the qTA algorithm as given earlier. In other words, computationally, one can try to find a target for any temporal interval. So, the real question is whether in a particular language syllable-sized targets exist and whether they are synchronized with the syllable. Evidence for synchronization of underlying tone with the entire syllable as opposed to only with the rhyme or the nucleus has been reported for Mandarin [15,18] and Cantonese [26]. It is shown in [27] that F_0 of unstressed syllables between surrounding stressed ones cannot be accounted for by linear or sagging interpolation between the F_0 peaks. The findings of consistent alignment of F_0 turning points in several non-tonal languages can be also viewed as evidence for target-syllable synchronization in these languages. More directly relevant here, there is some evidence that models that have pitch representation for each syllable tend to generate more natural-sounding prosody than those that either ignore the syllable or only have representations for the accented syllables [28-29]. Nevertheless, the issue of syllable synchronization is certainly unresolved, and much further research is needed.

3. Communicative Functions

It is the dream of prosody modeling to generate surface acoustic forms as close to those of natural production as possible. Therefore it may seem desirable to mark up all the prominent acoustic events in the training corpora so that models can be trained to reproduce them. This is exactly what we find in most of the annotation systems, including, in particular, ToBI [30], which, since its introduction in 1992, has become a widely-used for English, and it has been extended to many other languages, including Chinese [31], Japanese [32], Korean [33] . There are two problems with this popular approach, however. First, modeling is a process of predicting the acoustic output from input that is linguistically meaningful but phonetically abstract. Using predictors that already specify the output form makes the process potentially circular. One could argue, of course, that annotating vowels or lexical tones is also circular, as their identity already implies their surface form, especially in the case of tone. But this observation actually leads to the second problem, i.e., events that appear prominent in the F_0 track may not correspond to linguistically meaningful units. Note that in the case of vowels and tones, their identity are first and foremost determined by knowing that they distinguish words. For non-lexical prosodic components that are non-lexical, it is much trickier to determine what is linguistically meaningful.

One important reason why ToBI labeled events may not be linguistically relevant is that different communicative functions that are conveyed side by side may have simultaneously contributed to the surface prosodic events. thus an F_0 peak may contain lexical, focal and modality information at the same time [27,34]. This issue has been addressed by models that are superpositional, including the Fujisaki model [5,13], the SFC model [1] and the linear alignment model [8]. The advantages of the superpositional models can already be seen in these studies. What our approach differs from the superpositional models is that, although we also recognize simultaneous occurrence of multiple functions, we believe that their joint encoding is more complex than a uniform logarithmic addition algorithm.

Figure 3 shows a sketch of the Parallel Encoding and Target Approximation model (PENTA) proposed in [11]. The stacked boxes on the far left represent individual communicative functions as the driving force of the model. These communicative functions are realized through distinct encoding schemes (2nd stack of boxes from left) that specify the parameters of the articulatory process of target approximation (middle block). These parameters are then used to control the target approximation process to generate the acoustic output (right). With qTA described earlier, PENTA can be quantitative implemented and initial testing has yielded encouraging results [20]. In PENTA, therefore, all communicative functions contribute to the surface prosody by changing the TA parameters. But the nature and magnitude of the changes differ from function to function, and the patterns of these changes can be discovered through both empirical studies and modeling simulation.

Importantly, although this is an implementational assumption as explained earlier, each and every syllable is obligatorily assigned a target by the lexical function, and higher level functions are all encoded by making changes to the local functions. In cases where the changes introduced by a higher level function are mainly in terms of target height (*b*), the effect is likely similar to superpositional addition. But in many other cases, more complex changes are involved. For example, in Mandairn, focus expands the pitch range of the focused item, which involves raising *b* for the H tone,

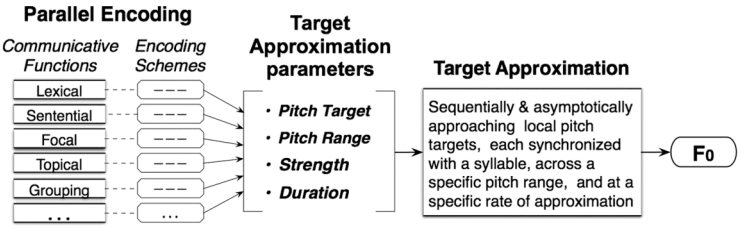

Figure 3. A sketch of the PENTA model. Adapted from [11]

lowering *b* for the L tone, and increasing the absolute value of *m* for the R and F tones. As we found in [20], similar changes in *b* and *m* are also introduced by focus in English.

The results of [20] suggest that training target approximation parameters based on communicative functions can effectively and efficiently generate surface prosody that sounds intelligible and natural. Nevertheless, further development of automatic procedures for function learning are needed before the system becomes usable in real applications.

4. Conclusions

The development of prosody models is crucial for the applicability of speech technology. In the last two decades, despite a drastic development of the speech science and technology, the use of prosody in practical systems is still limited. To enhance the applicability of prosody models, a major improvement is needed in the representation of prosody. In particular, it is necessary to adequately represent both the articulatory process of F_0 control and the communicative functions that make up the meanings of prosody. Various models have implemented either kind of representation in one way or another. But only PENTA and qTA — its computational realization have incorporated representations of both articulatory dynamics and communicative functions in a coherent way. While the theoretical benefits of such an integrated approach have been demonstrated, its practical advantage needs to be examined in further research.

5. References

[1] Bailly, G. and Holm, B., "SFC: A trainable prosodic model", Speech Commun., 46:348-364, 2005.

[2] Hirst, D. and Espesser, R., "Automatic modelling of fundamental frequency using a quadratic spline function", Travaux de l'Institut de Phonétique d'Aix, 15: 75-85, 1993.

[3] Pierrehumbert, J., "Synthesizing intonation", J. Acoust. Soc. Am., 70: 985-995, 1981

[4] Taylor, P., "Analysis and synthesis of intonation using the Tilt model", J. Acoust. Soc. Am., 107: 1697-1714, 2000.

[5] Fujisaki, H., "Dynamic characteristics of voice fundamental frequency in speech and singing", in P. F. MacNeilage [Ed], The Production of Speech, 39-55, New York: Springer-Verlag, 1983.

[6] Kochanski, G. and Shih, C., "Prosody modeling with soft templates", Speech Commun., 39:311-352, 2003.

[7] Hirschberg, J., "Communication and prosody: Functional aspects of prosody", Speech Commun., 36:31-43, 2002.

[8] van Santen, J. P. H. and Möbius, B., "A quantitative model of f0 generation and alignment", In A. Botinis [Ed], Intonation: Analysis, Modelling and Technology, 269-288, Kluwer Academic Publishers, 2000.

[9] Hirst, D. J., "Form and function in the represen-tation of speech prosody", Speech Commun., 46:334-347, 2005.

[10] Kohler, K., "Timing and Communicative Func-tions of Pitch Contours", Phonetica, 62:88-105, 2005.

[11] Xu, Y., "Speech melody as articulatorily imple-mented communicative functions", Speech Commun., 46:220-251, 2005.

[12] Kochanski, G., Shih, C. and Jing, H. "Quantitative measurement of prosodic strength in Mandarin", Speech Commun. 41:625-645, 2003.

[13] Fujisaki, H., Wang, C., Ohno, S. and Gu, W. "Analysis and synthesis of fundamental frequency contours of standard Chinese using the command-response model", Speech Commun., 47:59-70, 2005.

[14] Xu, Y., "Contextual tonal variations in Mandarin", J. Phonetics, 25:61-83, 1997.

[15] Xu, Y. "Consistency of tone-syllable alignment across different syllable structures and speaking rates", Phonetica, 55:179–203, 1998.

[16] Xu, Y., and Wang, Q. E., "Pitch targets and their realization: Evidence from Mandarin Chinese", Speech Commun., 33:319-337, 2001.

[17] Xu, Y., "Effects of tone and focus on the formation and alignment of F0 contours", J. Phonetics, 27:55-105, 1999.

[18] Xu, Y. "Fundamental frequency peak delay in Mandarin," Phonetica, 58:26-52, 2001.

[19] Xu, Y. and Sun, X., "Maximum speed of pitch change and how it may relate to speech", J. Acoust. Soc. Am. 111: 1399-1413, 2002.

[20] Prom-on, S., Xu, Y. and Thipakorn, B., "Modeling tone and intonation in Mandarin and English as a process of target approximation", J. Acoust. Soc. Am., 125:405-424, 2009.

[21] Gribble, P. L., Mullin, L. L., Cothros, N. and Mattar, A., "Role of cocontraction in arm movement accuracy", J. Neurophysiol., 89:2396-2405, 2003.

[22] Fowler, C. A., "Coarticulation and theories of extrinsic timing," J. Phonetics 8:113-133, 1980.

[23] Pierrehumbert, J., "The Phonology and Phonetics of English Intonation," Ph.D. dissertation, MIT, Cambridge, MA, 1980.

[24] Saltzman, E. L., and Munhall, K. G., "A dynamical approach to gestural patterning in speech production," Ecological Psychology 1:333-382, 1989.

[25] Chen, Y., and Xu, Y., "Production of weak elements in speech — Evidence from f0 patterns of neutral tone in standard Chinese," Phonetica 63:47-75, 2006.

[26] Wong, Y. W., and Xu, Y., "Consonantal perturbation of f0 contours of Cantonese tones," in Proceedings of The 16th International Congress of Phonetic Sciences, Saarbrucken, 1293-1296, 2007.

[27] Xu, Y. and Xu, C. X., "Phonetic realization of focus in English declarative intonation" J. Phonetics, 33:159-197, 2005.

[28] Sun, X., "The determination, analysis, and synthesis of fundamental frequency," Ph.D. dissertation, Northwestern University, 2002.

[29] Raidt, S., Bailly, G., Holm, B. *et al.*, "Automatic generation of prosody: Comparing two superpositional systems," in Proceedings of Speech Prosody 2004, Nara, Japan, 417-420, 2004.

[30] Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P. and Hirschberg, J., "ToBI: A Standard Scheme for Labeling Prosody", in Proc. ICSLP 1992, 867-869.

[31] Aijun, L., "Chinese prosody and prosodic labeling of spontaneous speech", in Proc. Speech Prosody 2002, 39-46.

[32] Venditti, J., "Discourse structure and attentional salience effects on Japanese intonation", PhD Thesis, Ohio State University, 2000.

[33] Jun, S., "K-ToBI (Korean ToBI) Labeling Conventions", Speech Science, 7:143-170.

[34] Liu, F. and Xu, Y., "Question intonation as affected by word stress and focus in English", in Proc. The 16th ICPhS, Saarbrücken: 1189-1192.