Modeling Speech Melody as Communicative Functions with PENTAtrainer2

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

[1]Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand
[2]Department of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom

santitham@cpe.kmutt.ac.th, yi.xu@ucl.ac.uk

Abstract

This paper presents PENTAtrainer2, a semi-automatic software package written as Praat plug-in integrated with Java programs, and its applications for analysis and synthesis of speech melody as communicative functions. Its core concepts are based on the Parallel Encoding and Target Approximation (PENTA) framework, the quantitative Target Approximation (qTA) model, and the simulated annealing optimization. This integration allows it to globally optimize for underlying pitch targets of specified communicative functions. PENTAtrainer2 consists of three computational tools: Annotation tool for defining communicative functions as parallel layers, Learning tool for globally optimizing pitch target parameters, and Synthesis tool for generating speech melody according to the learned pitch targets. Being both theory-based and trainable, PENTAtrainer2 can serve as an effective tool for basic research in speech prosody.

Index Terms: prosody modeling, parallel encoding, target approximation, communicative function, stochastic optimization

1. Introduction

Speech prosody conveys communicative meanings through the manipulation of fundamental frequency (F_0), duration, intensity, and voice quality. Of these cues, F_0 is one of the most important. Communicative function refers to the relation between a specific communicative meaning and how it is encoded in the structure of speech melody. Modeling communicative function is thus a key to achieve effective prosody analysis and synthesis.

This paper presents PENTAtrainer2, a tool for prosody analysis and synthesis. It was created with an ultimate goal to assist speech researchers in prosody modeling studies. It provides users easy-to-use interfaces to perform three critical tasks in prosody modeling: data annotation, parameter estimation, and prosodic prediction. The program and the step-by-step tutorials in prosody analysis and synthesis can be freely downloaded from (http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer2/).

2. PENTAtrainer2

PENTAtrainer2 (pen·ta·train·ner·two) consists of a set of Praat scripts that facilitate the investigation of underlying representations of communicative functions in any language [1]. Its core concept is based on the Parallel Encoding and Target Approximation (PENTA) framework [2]. PENTAtrainer2 encapsulates the quantitative Target Approximation (qTA) model, which represents dynamic F0 control [3], and simulated annealing optimization [4], which is a stochastic learning algorithm used to globally optimize model parameters. Provided with annotated sound files, PENTAtrainer2 automatically learns the optimal parameters of all possible functional combinations that users have annotated. After the optimization, the learned functional parameters can be used to synthesize F_0 contours according to any of the given communicative functions. Summaries of the modeling technique will be briefly discussed in the following sub-section.

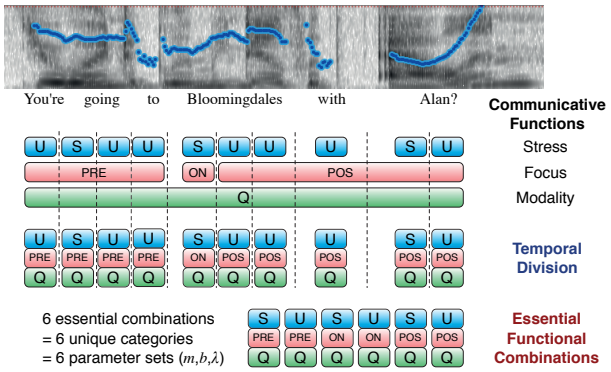

Figure 1: *An illustration of the conversion from the parallel functional annotation to the essential functional combinations.*

2.1. Parallel annotation of communicative functions

PENTAtrainer2 is a data-driven prosody modeling software. The specific values of the model parameters are optimized from the training speech material. The basic idea is to identify the number of functional layers and their corresponding prosodic categories that span across specified temporal units (e.g. tone localized with the syllable). It is critical for the system to know what to learn. Fig. 1 illustrates the annotation of three communicative functions of English intonation: Stress, Focus, and Modality. Each layer was annotated independently and the function-internal categories are defined manually by the investigator. Boundaries on each layer were marked according to the time span of that prosodic event, again defined by the investigator. For example, in Fig. 1, the "Stress" layer is associated with the syllable and can have two values: Stressed (S) and Unstressed (U). For a "Focus" layer, PRE, ON, POS denote pre-focus, on-focus, and post-focus regions respectively. For a "Modality" layer, Q denotes question and S denotes statement. Note that the names here carry no meaning to PENTAtrainer2, as all it cares is which are the same categories and so should be given a common set of target parameters. This differs from annotation schemes in which the names are meaningful (e.g., ToBI [5], INTSINT [6]).

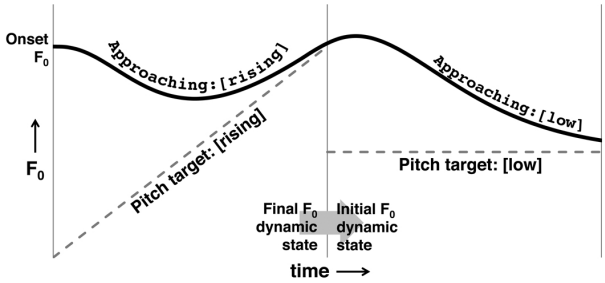

Figure 2: *Illustration of target approximation process [3, 7].*

2.2. Modeling F_0 movement with qTA model

To model F_0 movement, PENTAtrainer2 uses the quantitative Target Approximation (qTA) Model [3], which is based on the theoretical target approximation model [7]. Fig. 2 is an illustration of the basic concept of target approximation. Surface F_0 contours (solid curve) are the responses of the target approximation process to the driving force of pitch targets (dashed lines). These targets represent the goals of the F_0 movement and are synchronized to the host syllable. Pitch targets are sequentially implemented syllable by syllable, starting from the beginning of the utterance. At the boundary of two syllables, the F_0 dynamic state at the end of the preceding syllable is transferred to the next syllable.

In qTA, a pitch target is defined as a forcing function that drives the F_0 movement. It is mathematically 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. t is a relative time from the syllable onset. The F_0 control is implemented by 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 of the system. 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 specified segment. The parameter λ represents the strength of the target approximation movement. The initial F_0 dynamic state consists of initial F_0 level, $f_0(0)$, velocity $f_0'(0)$, and acceleration, $f_0''(0)$. The dynamic state is transferred from one syllable to the next at the syllable boundary to ensure continuity of F_0. The three transient coefficients are computed with the following formulae.

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

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

$$c_3 = (f_0''(0) + 2c_2\lambda \quad c_1\lambda)/2 \tag{5}$$

qTA thus defines each pitch target with only three parameters, m, b, and λ. Of the three parameters, m and b are used to specify the form of the pitch target. For example, the Mandarin Rising and Falling tones, which differ mainly in target slope, have positive and negative m values, respectively; the Mandarin High and Low tones, which differ mainly in target height, have relatively high and low b values, respectively [3, 8, 9]. Here the value of b is relative to a reference pitch, which can be either the speaker F_0 mean or the initial F_0 of an utterance. λ specifies how rapidly the target is approached, with a larger value indicating faster approximation. This approximation rate can define an additional property of a tone. For example, the Mandarin neutral tone is found to have a much smaller λ value than the full tones [9], which is consistent with the observation that the neutral tone may have a weak articulatory strength [10].

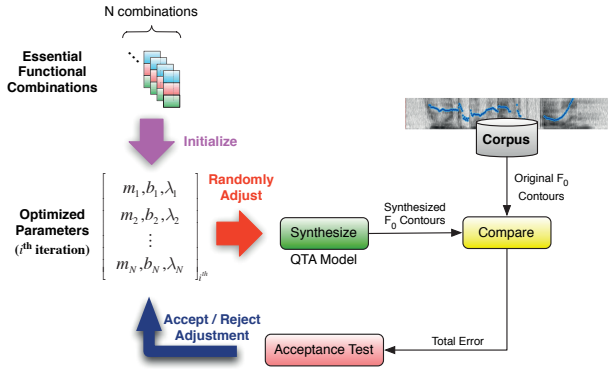

Figure 3: *A diagram illustrating the application of the simulated annealing algorithm used for globally optimizing parameters of essential functional combinations.*

2.3. Parameter optimization

In PENTAtrainer2, the parameter estimation is done via a stochastic global optimization that can directly estimate parameters of functional categories in a corpus. The general idea is illustrated as a block diagram in Fig. 3. At the initial stage, the algorithm randomly modifies parameters of all functional categories and tests whether or not such modification is acceptable by a probabilistic method. The number of initialized parameter sets is equal to the number of essential functional combinations obtained from the procedure to be discussed in the next section. These parameters are randomly adjusted and used in qTA to synthesize F0 contours which are compared to the original data. The total sum of square error between original and synthesized F0 contours calculated from the whole corpus is then used to determine whether the proposed adjustment is acceptable. The decision to accept or reject the proposed adjustment depends on the acceptance probability calculated from the change in error incurred from parameter adjustment and the annealing temperature,

$$p_{th} = \exp((E_{current} \quad E_{previous})/T) \tag{6}$$

where $E_{current}$ and $E_{previous}$ are the total sum of square errors calculated from the whole corpus. The difference between these two errors indicates the change in the total error incurred from the parameter adjustment. T is the annealing temperature which controls the degree at which a bad solution is allowed. In the decision process, a random testing probability p_{test} is generated and compared to p_{th}. If $p_{test} < p_{accept}$, the parameter adjustment is accepted; otherwise it is rejected. T is initially set to a high value and then gradually reduced as the procedure is repeated. This allows the solution to converge close to the global optimum over iterations.

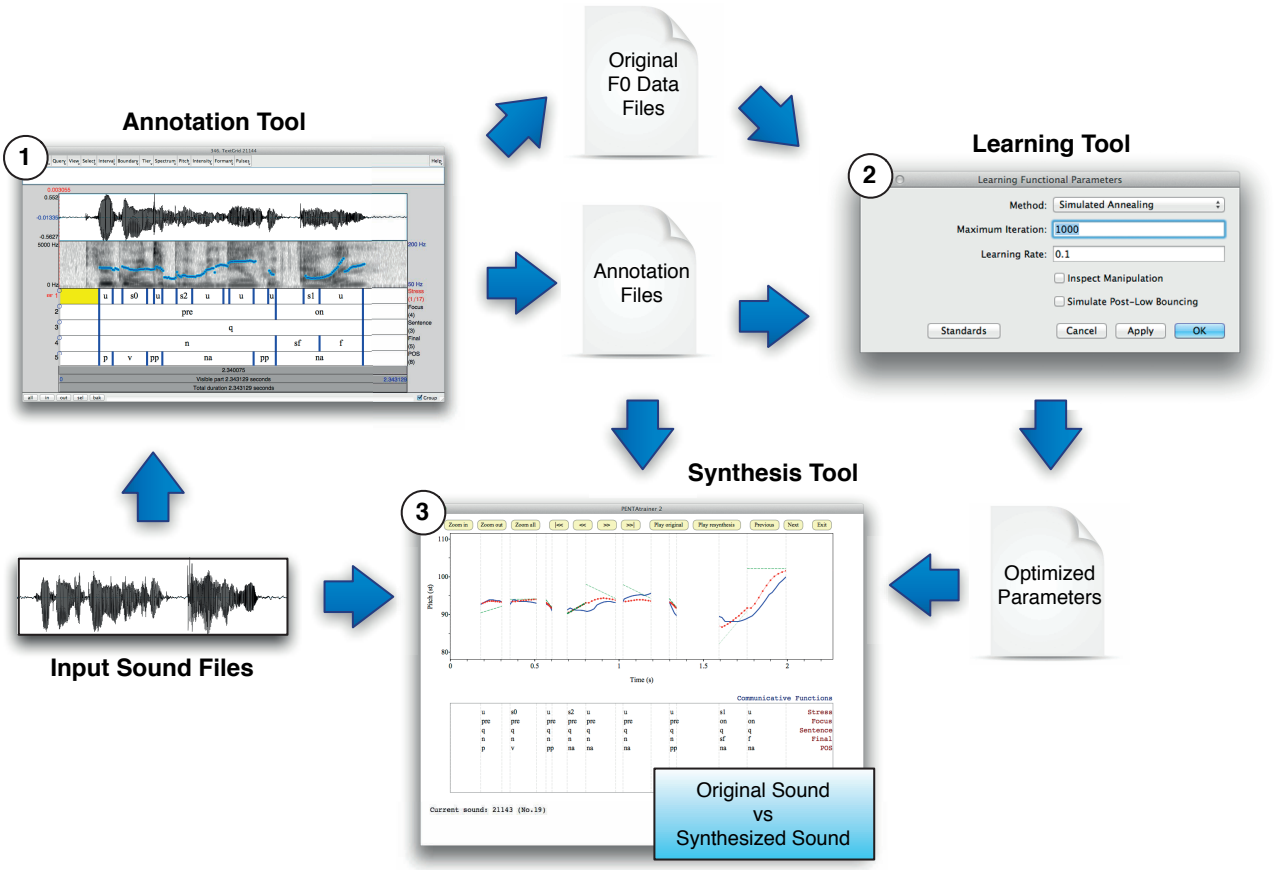

Figure 4: *A workflow of analysis and synthesis of speech melody using PENTAtrainer2.*

2.4. Prosody modeling workflow

PENTAtrainer2 composes of three main tools: Annotate, Learn, and Synthesize. Each of them can be accessed individually in the PENTAtrainer2 plug-in menu. Each tool corresponds to a main task in the prosody modeling workflow shown in Fig. 4. First, the speech corpus is annotated using the Annotate tool. Communicative functions related to a corpus are annotated in separate tiers. Two tiers, tone and vowel length, are annotated for the Thai corpus in this project. Temporal boundaries in each tier are aligned consistently to the prosodic or segmental events of that tier. For the present study, because both tone and vowel length boundaries are synchronized to the syllable, the syllable boundaries are used as temporal markings for both tiers. The co-occurrences of events in the two tiers form functional combinations, which represent interactions between tiers. The annotation step is done iteratively for each sound file in the corpus. In this step, investigators can also inspect and manually rectify the vocal pulse marks used in F0 calculation. This step requires the most human effort. Next, after all the sound files are annotated, the second step is to estimate the pitch target parameters using the Learn tool. Investigators only need to provide the Learn tool with the optimization parameters, and it will then automatically estimate the optimal parameters of the functional combinations. The third and the last step allows investigators to synthesize or predict the F0 contours from the learned parameters (or humanly provided parameters if so desired) using the Synthesize tool. The optimized parameters can be either speaker-dependent, i.e., learned from each individual speaker, or speaker-independent, i.e., derived by averaging the parameters of all the speakers.

3. Conclusion

This paper presents the technical detail of PENTAtrainer2, and its workflow for prosody modeling. It provides analysis and synthesis functionalities to represent speech prosody as communicative functions. It has been found to be effective in capturing underlying representation of communicative functions in several languages and able to synthesize with high accuracy [1]. Being both theory-based and trainable, PENTAtrainer2 can serve as a modeling tool for basic research in speech prosody.

4. Acknowledgement

The authors would like to thank for the financial supports the Royal Society (UK) and the Royal Academy of Engineering (UK) through the Newton International Fellowship Scheme, the Thai Research Fund through the TRF Grant for New Researcher, and the National Science Foundation.

5. References

[1] Xu, Y. and Prom-on, S., "From variable surface contours to invariant underlying representations: Synthesizing speech melody via model-based stochastic learning", Manuscript submitted for publication, 2013.

[2] Xu, Y., Speech melody as articulatory implemented communicative functions, Speech Commun., 46(3-4): 220-251, 2005.

[3] 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.

[4] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P., Optimization by simulated annealing, Science, 220(4598): 671-680, 1983.

[5] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J., ToBI: A standard for labeling English prosody, In: Proc. ICSLP 1992, Banff, pp. 867-870, 1992.

[6] Hirst, D. J., The analysis by synthesis of speech melody: From data to models, J. Speech Science, 1:55-83, 2011.

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

[8] Prom-on, S., Liu, F., and Xu, Y., Functional modeling of tone, focus, and sentence type in Mandarin Chinese, In:Proc. ICPhS XVII, Hong Kong, pp. 1638-1641, 2011.

[9] Prom-on, S., Liu, F., and Xu, Y., Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling J. Acoust. Soc. Am., 132: 421-432, 2012.

[10] 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.