PENTATrainer2: A hypothesis-driven prosody modeling tool

Santitham Prom-on, Yi Xu

Dept of Speech, Hearing, and Phonetic Sciences, University College London, UK

Abstract

Prosody is an essential aspect of speech, as it carries both lexical and non-lexical information. A conventional approach to studying speech prosody is to collect and analyze F_0 data based on certain hypotheses and then develop a theory based on the observation, which constitutes the final conclusion of the study. This process is however far from complete, as the developed theory has not been actually tested for its ability to predict actual acoustic data. This paper presents PENTATrainer2, a prosody modeling tool based on the parallel encoding and target approximation (PENTA) framework. PENTATrainer2 can facilitate prosody studies in testing hypotheses and theories using an automatic analysis-by-synthesis and stochastic learning algorithm. Users can flexibly design the annotation scheme based on their own hypotheses and then find out whether the hypothesized categories can lead to accurate synthetic F_0 contours. PENTATrainer2 consists of three main components: multi-layer annotation, target approximation and stochastic optimization. First, acoustic data are annotated in parallel layers, each of which corresponds to a functional category that may affect F_0 contours. These layers are then compiled into unique functional combinations. The combinations represent underlying invariant representations of communicative functions and their interaction with each other. Target approximation parameters of each combination are then learned through analysis-by-synthesis and stochastic optimization. Pilot tests of PENTATrainer2 conducted on Thai, Mandarin and English demonstrate not only high accuracy of the synthesized F_0 contours but also distinctive contrasts in the distribution of pitch target parameters. This indicates the effectiveness of PENTATrainer2 in modeling speech prosody.

Keywords: prosody modeling, analysis-by-synthesis, parallel encoding, target approximation, stochastic optimization.

Introduction

Speech prosody conveys multiple levels of information simultaneously, in terms of both linguistic contrasts such as tone, focus and modality, and paralinguistic variations related to emotion, mood and attitude. Usually, the method of studying prosody is to try to link such information to changes in surface acoustics by means of statistical analysis. A conclusion drawn from the results was then used to formulate a theory about prosody. This process is however far from complete, as the developed theory has not been actually tested for its ability to predict actual acoustic data. This is a crucial step as it makes the formulated theory testable. A major reason for the general absence of this step is the lack of quantitative tools that allow speech scientists to incorporate their empirical findings into quantitative modeling. The present paper reports the development of PENTATrainer2 as one of such tools, which can automatically learn parameters of user-defined prosodic categories and synthesize F_0 contours according to the learned parameters.

PENTATrainer2

Modeling Method

The general scheme of PENTAtrainer2 is based on the notion that speech prosody conveys information about multiple communicative functions in parallel (Xu, 2005). This notion is implemented in PENTAtrainer2 in its annotation scheme. Figure 1 shows an example of parallel annotation scheme of three communicative functions of English intonation, including Stress, Focus, and Modality. Each function was annotated independently as a parallel layer. Boundaries on each layer were marked according to the time span of that function.

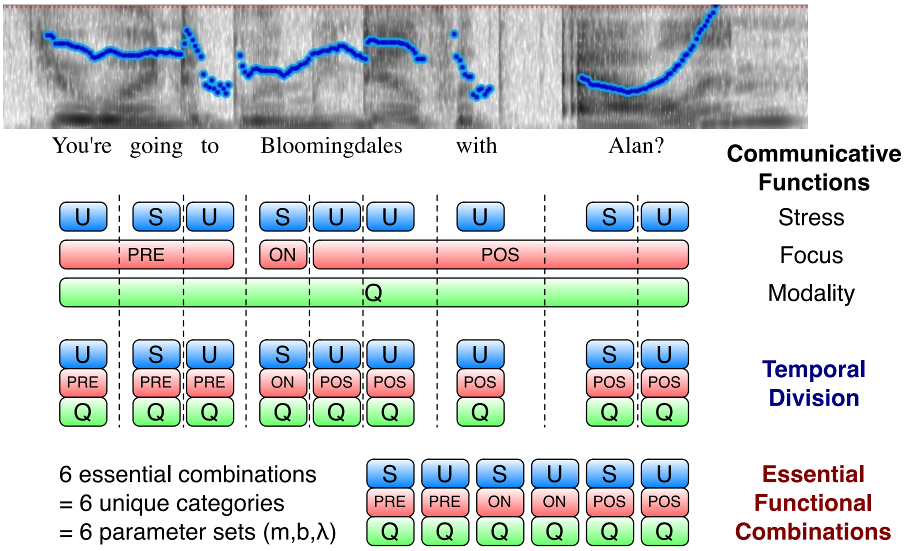


Figure 1. An example of conversion process from parallel annotations to essential functional combinations.

These parallel layers of communicative functions can also be considered as a sequence of functional combination categories. By projecting the boundaries from the layer with the smallest temporal unit (i.e. largest number of intervals) to other layers, we can obtain a sequence of functional combinations associating with each interval. Summarizing the unique combinations of all utterances in the corpus together results in a set of functional combinations that essentially describe the prosody of that corpus.

Each interval, which is temporally divided from the functional combinations, corresponds to an F_0 movement that approaches one pitch target. This movement is quantitatively implemented in the quantitative Target Approximation (qTA) model (Prom-on et al., 2009). Figure 2 illustrates an example of F_0 movements and their corresponding pitch targets in the qTA model. In qTA, F_0 asymptotically approaches consecutive pitch targets and its dynamic states are transferred from one target approximation interval to the next at the boundary. This transfer of dynamic states, which include F_0 level, velocity, and acceleration, allows the process to carryover the momentum of the previous syllable, thus resulting in the observed carryover coarticulation. F_0 movement thus contains two components: forced response and transient response. Forced response is a pitch target, which is the goal driving the target approximation process, while transient response is the F_0 transition from the initial F_0 dynamic state to the pitch target.

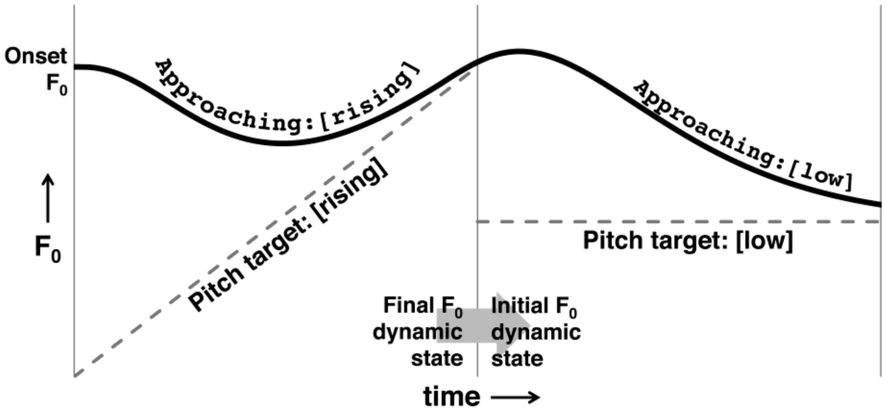


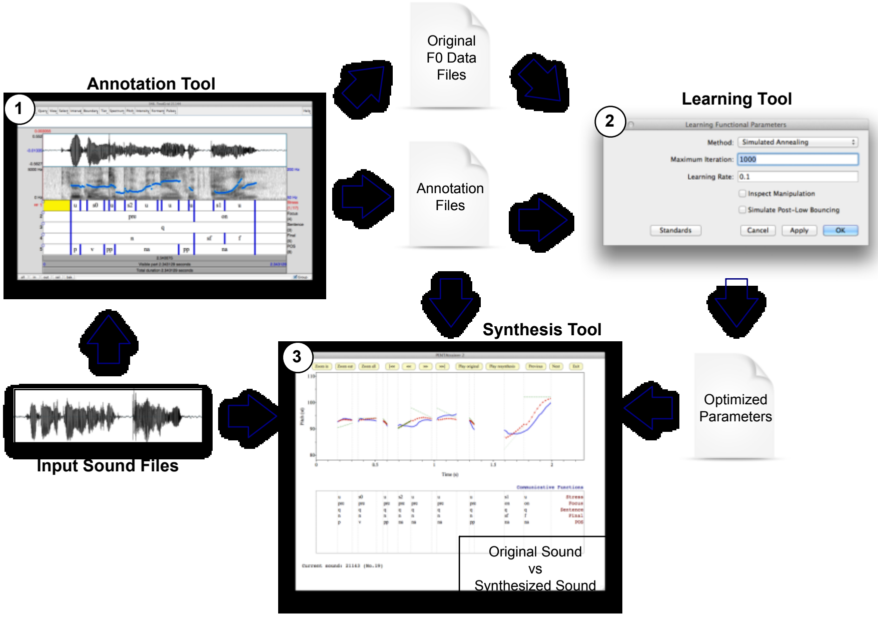
Figure 2. Illustration of the target approximation process (Xu and Wang, 2001; Prom-on et al., 2009).

In qTA, there are three model parameters controlling the F_0 movement of each interval, including target slope (m), target height (b), and strength of target approximation (λ). m and b specify the form of the pitch target and λ indicates how rapidly a pitch target is approached.

After the functional combinations were determined, their parameters were estimated using the analysis-by-synthesis strategy and the simulated annealing algorithm (Kirkpatrick et al., 1983). Parameters of essential combinations were randomly initialized. They were then randomly modified and tested to determine whether to accept or reject the proposed modification depending on the annealing temperature of the algorithm. The temperature is initially set to a high value and then gradually reduced as the procedure is repeated. This allows the solution to converge to the global optimum over the iterations. Since the final optimized parameters may differ slightly, the learning process should be repeated a number of times to obtain more stable solution. The median of the parameters were then calculated across repetitions for each functional category of each speaker.

Software

PENTATrainer2 contains three computational tools. Figure 3 shows the workflow of PENTATrainer2. First, users need to manually define the annotation scheme using the Annotation tool. Next, parameters are automatically optimized using the Learning tool. Users can then use the Synthesis tool to synthesize F_0 contours based on the optimized parameters and compare them to the original contours.



Learning Functional Parameters

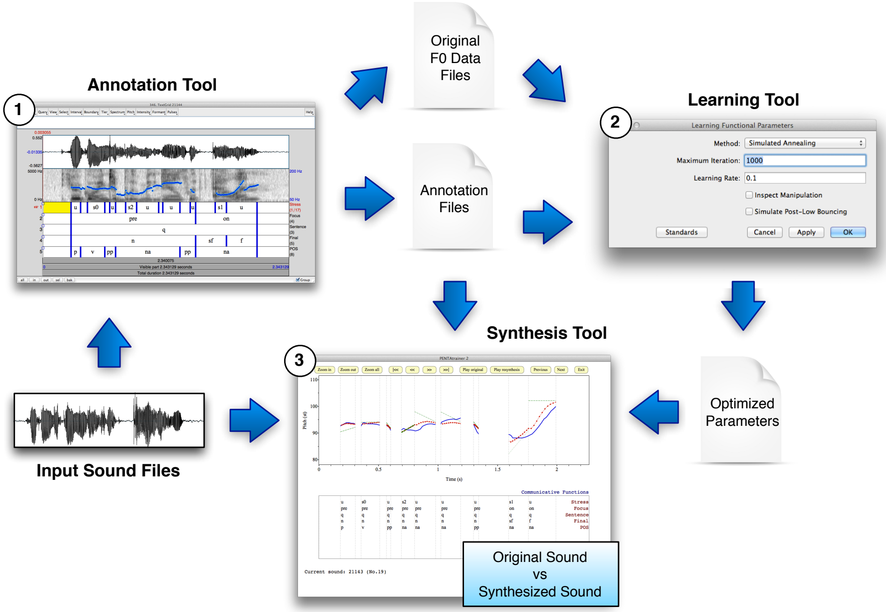
Method: Simulated Annealing
Maximum Iteration: 1000
Learning Rate: 0.1
□ Inspect Manipulation
□ Simulate Post-Low Bouncing

Figure 3. PENTATrainer2's workflow for prosody modeling

Pilot Tests

Corpora

We conducted pilot tests of PENTATrainer2 on Thai, Mandarin and English corpora. Table 1 shows the detail of the corpora. For full details of each corpus, please refer to prior publications (Thai: Prom-on and Xu, 2012; Mandarin: Prom-on et al., 2011; English: Liu and Xu, 2007). Each corpus was annotated separately according to the prosodic factors of that study.



Method: Simulated Annealing
Maximum Iteration: 1000
Learning Rate: 0.1
□ Inspect Manipulation
□ Simulate Post-Low Bouncing

Figure 3. PENTATrainer2’s workflow for prosody modeling

Parameters were estimated according to the method described in Section 2.1. Root-mean-square error (RMSE) and Pearson's correlation coefficient were used to measure the synthesis accuracy between the synthesized and original F_0 contours.

Table 1. Corpus descriptions

Corpora	Number of Utterances (Subjects)	Factors
Thai	2500 (3 males, 2 females)	Tone, Vowel Length
Mandarin	1280 (4 males, 4 females)	Tone, Focus, Modality
English	960 (2 males, 3 females)	Stress, Focus, Modality

Results

Figure 4 shows examples of synthesized F_0 contours of all three languages as compared to the original F_0 contours. As can be seen, the F_0 contours synthesized with learned categorical pitch targets are very close to the original ones. Table 2 shows the overall synthesis accuracies of all three corpora. These accuracies are better than when parameters were estimated locally (Prom-on et al., 2009, 2011). High correlations and relatively low RMSEs can be seen across languages. Such high synthesis accuracies provide support for the user-defined hypothesized functional categories. These results also indicate the effectiveness and the generalizability of PENTATrainer2 to different languages.

Table 2. Means and standard errors of RMSE and correlation of each corpus. Parameters were learned according to the factors shown in Table 1.

Corpora	Number of Parameters	RMSE (semitone)	Correlation
Thai	10/subject	0.78 ± 0.05	0.889 ± 0.012
Mandarin	28/subject	2.16 ± 0.22	0.903 ± 0.008
English	26/subject	2.07 ± 0.23	0.836 ± 0.019

After obtaining the parameters that yield the best synthesis accuracy, the next step in a general modeling study is to analyze the distribution of estimated parameters to determine whether there is any difference between categories. This can lead to a better understanding of the underlying representations of that prosodic phenomenon. To demonstrate this, we show here the parameter distributions of Thai tones and their related statistical analyses.

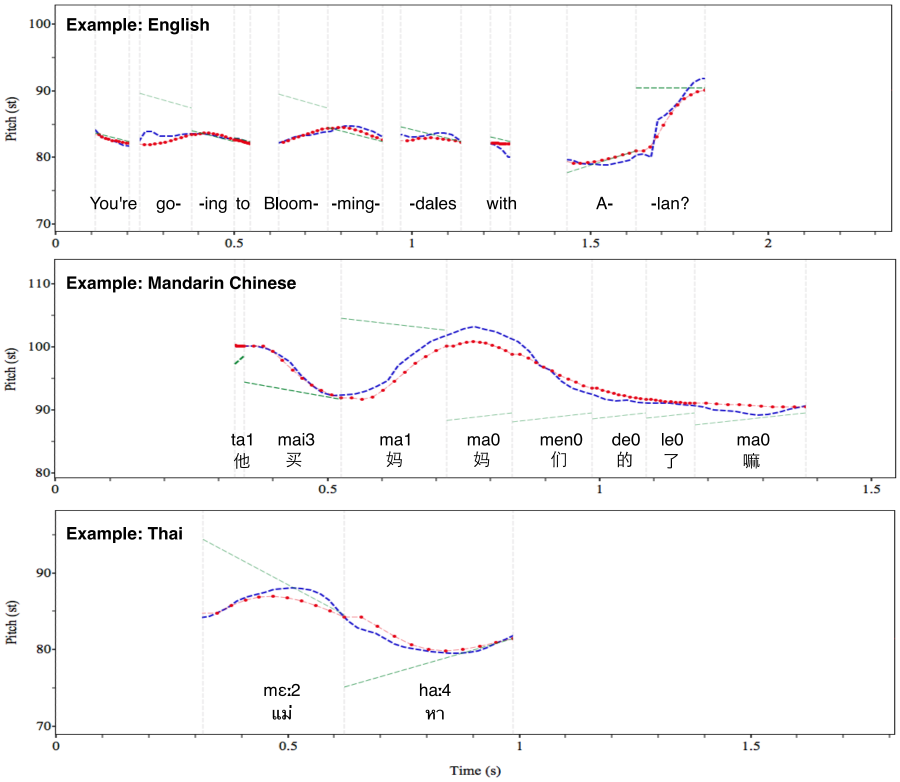


Figure 4. Examples of synthesized (red dotted line) compared to the original (blue dashed line) F_0 contours. The short dashed green lines represent the learned categorical pitch targets with which the synthetic F_0 contours were generated.

Thai has five lexical tones, including Mid (M), Low (L), Falling (F), High (H) and Rising (R), and two lexical vowel lengths, short and long. Previous acoustic analysis has shown highly variable F_0 contours of these tones in connected speech depending on both contexts and other lexical factors such as vowel length. In particular, there are both carryover and anticipatory effects in contextual tonal variations (Gandour *et al.*, 1992; Potisuk *et al.*, 1997). There are also interactions between tone and vowel length (Gandour, 1977), with the shorter duration associated with higher F_0 value. But it is unknown whether these variations reflect changes in the underlying tonal representation. In this study, we explored these issues by making use of PENTATrainer2's ability to learn underlying representations. Tone and vowel length were labeled without contextual information before the training process. Figure 5 shows the distributions of pitch target parameters learned using PENTATrainer2. Repeated measures ANOVAs showed that the parameters were significantly different depending on the tonal categories (*m*: F(4,49) = 56.81, $p < 0.001$; *b*: F(4,49) = 71.07, $p < 0.001$; λ: F(4,49) = 9.23, $p < 0.001$). This indicates that the variability within tone groups is significantly less than that between groups. This also indicates that despite the variability in surface acoustics, learned underlying tonal representations are consistent and can accurately simulate F_0 contours that varied depending on the tonal contexts.

Comparing between different vowel lengths, target slope and strength were not significantly different, but target height of M tone was higher in short vowels than in long vowels (F(1,49) = 5.37, $p = 0.026$). This difference might suggest that M has two tonal targets so as to enhance the vowel length contrast similar to what is found in Finnish (Vainio *et al.*, 2010). However, we cannot reach a clear conclusion on this because the difference in the learned target height could also be due to other factors. For example, it is possible that the height difference is because M has a weak strength, just like the Mandarin neutral tone (Chen and Xu, 2006). But the estimation of such weak strength requires the presence of consecutive M tones (Prom-on et al., 2012), which is lacking in the current corpus. This issue therefore has to be resolved by future studies.

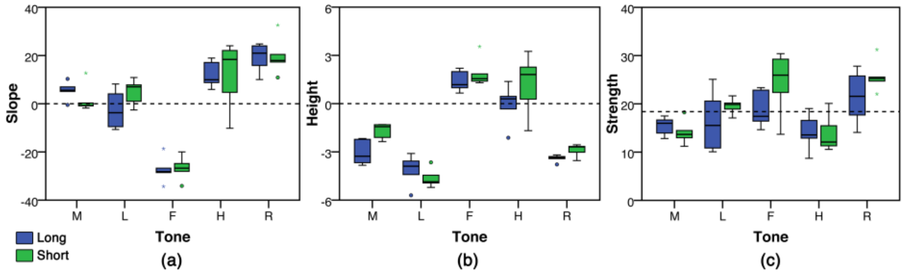


Figure 5. Parameter distributions of Thai tones (Prom-on and Xu, 2012).

Conclusion

This paper presents PENTATrainer2 and its workflow for prosody modeling. PENTATrainer2 can learn underlying representations of communicative functions in the form of pitch target parameters, and use them to accurately synthesize F_0 contours. Users can flexibly design hypothesized functional categories and test whether they lead to an improvement in synthesis accuracy. This allows speech scientists to objectively and quantitatively investigate speech prosody based on communicative functions. The pilot test results have provided initial indication that PENTATrainer2 works effectively across languages. Both high synthesis quality and its ability to estimate underlying representations indicate the effectiveness of PENTATrainer2 in prosody modeling. The integration of the analysis-by-synthesis approach and the stochastic optimization also allows users to explore theoretical issues such as underlying representations of tonal and intonational units.

Acknowledgements

We would like to thank the Royal Academy of Engineering for the financial support through the Newton International Fellowship Scheme. We also thank Fang Liu for providing the English and Mandarin Chinese corpora used in this work.

References

Chen, Y. and Xu, Y. (2006). Production of weak elements in speech -- Evidence from f0 patterns of neutral tone in standard Chinese. *Phonetica* **63**: 47-75.

Gandour, J. 1977. On the interaction between tone and vowel length: Evidence from Thai dialects. Phonetica 34, 54-65.

Gandour, J., Potisuk, S., Dechongkit, S., and Ponglorpisit, S. 1992. Tonal coarticulation in Thai disyllabic utterances: a preliminary study. Linguistics of the Tibeto-Burman Area 15, 93-110.

Liu, F., and Xu, Y. 2007. Question intonation as affected by word stress and focus in English. Proc. ICPhS 2007, 1189-1192.

Potisuk, S., Gandour, J., and Harper, M. P. 1997. Contextual variations in trisyllabic sequences of Thai tones. Phonetica 54, 22-42.

Prom-on, S., Liu F., and Xu, Y., 2011. Functional modeling of tone, focus and sentence type in mandarin Chinese. Proc. ICPhS 2011, 1638-1641.

Prom-on, S., Liu, F. and Xu, Y., 2012. Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling. Journal of the Acoustical Society of America 132, 421-432.

Prom-on, S., and Xu, Y. 2012. Pitch target representation of Thai tones. Proceedings of TAL 2012, Nanjing, China.

Prom-on, S., Xu, Y., and Thipakorn, B., 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.

Vainio, M., Järvikivi, J., Aalto, D., and Suni, A. 2010. Phonetic tone signals phonological quantity and word structure. Journal of the Acoustical Society of America 128, 1313-1321.

Xu, Y., 2005. Speech melody as articulatorily implemented communicative functions. Speech Communication 46, 220-251.

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