Deriving underlying tonal representations through computational modeling

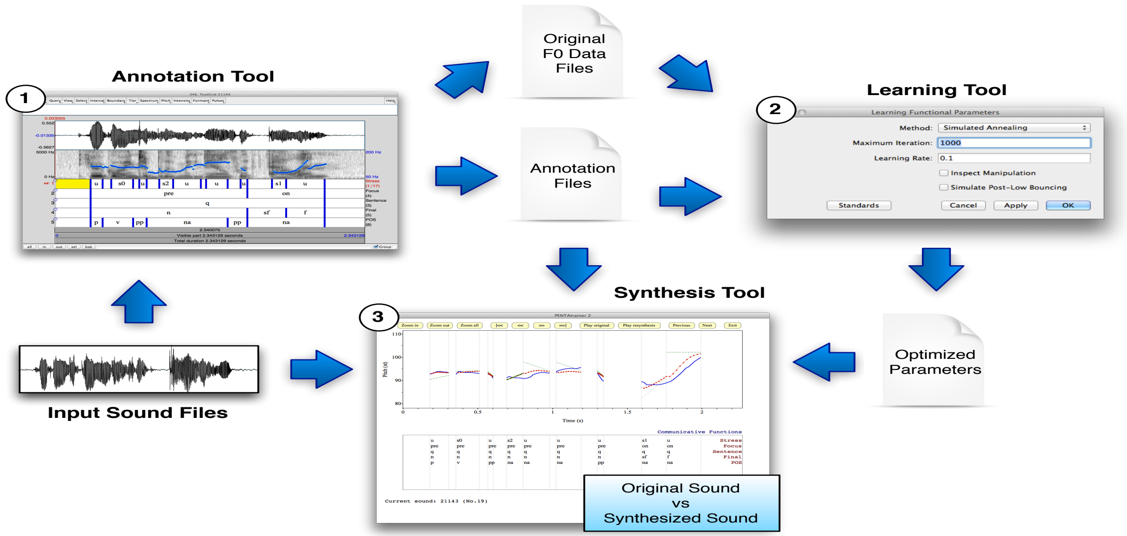
Santitham Prom-On and Yi Xu
University College London, UK; santitham.prom-on@ucl.ac.uk, yi.xu@ucl.ac.uk

Variations of fundamental frequency (F_0) in speech come not only from communicative functions such as tone, focus, and sentence modality, but also from physiological factors such as coarticulation. Conventional approaches to tone and intonation generally try to link communicative functions to F_0 changes directly, thus running the risk of confounding the underlying representations with effects of coarticulation.

In this paper, we show that it is possible to estimate underlying representations of tone and intonation free of coarticulatory confound. This is done with PENTATrainer2, a prosody modeling tool that combines built-in coarticulatory mechanisms, parallel encoding annotation and global stochastic optimization (Xu, 2005; Prom-on *et al*., 2009). Given user-defined temporal boundaries and categorical annotations, PENTAtrainter2 automatically learns invariant underlying pitch targets associated with user-postulated tonal and intonational functions using analysis-by-synthesis and simulated annealing. It also allows users to perform deductive hypothesis testing by designing hypothesis-specific annotation schemes. Figure 1 illustrates the workflow of PENTATrainer2.

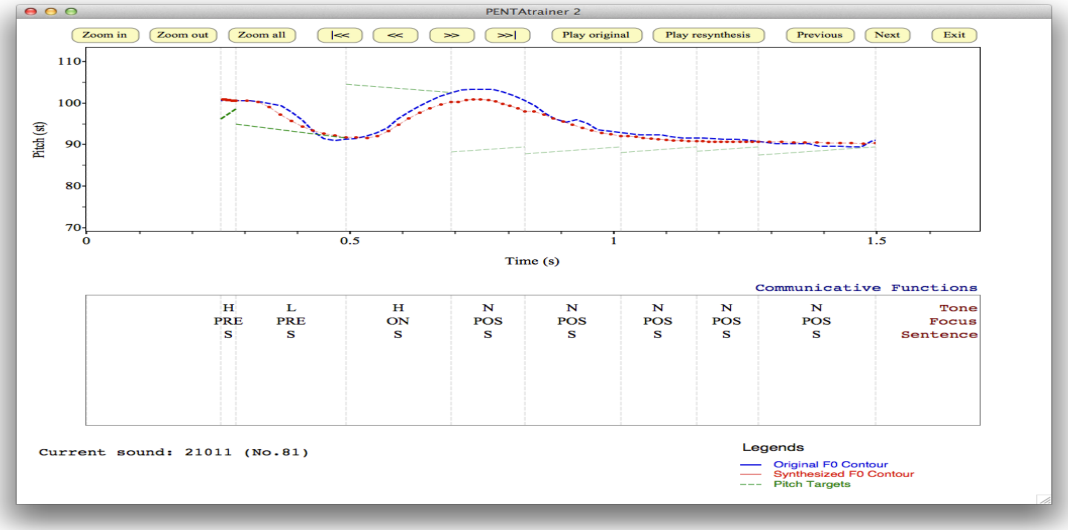
We applied PENTATrainer2 to a Mandarin corpus, consisting of 1280 eight-syllable utterances by 4 male and 4 female native speakers, to learn categorical parameters of tone, focus, and sentence modality. The F_0 contours synthesized with the learned parameters yielded average root-mean-square error (RMSE) of 2.16 semitones and average Pearson's correlation coefficient of 0.903. Figure 2 shows an example of the comparison between the original and synthesized F_0 contours generated by Synthesis tool in PENTAtrainer2. The local fit between the contours is very good, even though the parameters are highly categorical, having been optimized from all the utterance of the same speaker. Figure 2 also shows how the Mandarin Neutral (N) tone, which is known to be severely influenced by the preceding tones, can be represented by a mid target with weak approximation strength. This same set of parameters can be used to accurately synthesize F_0 contour of N tone in other preceding tonal contexts as well, hence eliminating the need to treat it as targetless (Shih, 1987) or underspecified (Myers, 1999).

To examine PENTATrainer2's ability to do hypothesis testing, we tested three hypotheses about the representation of the Mandarin Low (L, tone 3) tone in the case of tone sandhi, including (1) same as the no-sandhi variant, (2) same as the Rising (R, tone 2) tone, and (3) as a new tone category. The synthesis accuracies of case (2) and (3) are not significantly different (RMSE: $p = 0.414$; Correlation: $p = 0.394$) with case (3) slightly more accurate than case (2). Case (1) has significantly lower accuracy than both of the other two cases (1vs2; RMSE: $p < 0.001$; Correlation: $p = 0.002$; 1vs3; RMSE: $p < 0.001$; Correlation $p < 0.001$). These results indicate that the underlying target of L tone sandhi can be either a separate category or the same as the R-tone target, but clearly not a L-tone target, thus confirming previous findings based on acoustic analyses and perceptual tests. This case study indicates the potential of PENTATrainer2 in deductively testing theories based on its performance in synthesis.



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

Figure 1. A workflow of PENTATrainer2, which technically consists of Annotation, Learning, and Synthesis tools. The number in the circle on the top-left corner of each tool indicates the order of general steps for modeling speech prosody.



Current sound: 21011 (No.81)

Figure 2. A screenshot of the Synthesis tool comparing the original (blue) and synthesized (green) F_0 contours based on optimized parameters. Pitch target of each interval are display as a dashed green line whose thickness represents the strength of its target approximation.

References

Myers, S. (1998) Surface underspecification of tone in Chichewa. *Phonology* **15**: 367-392.

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

Shih, C. (1987) The phonetics of the Chinese tonal system, AT&T Bell Labs technical memo.

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