Predictive synthesis of Japanese word prosody using AMtrainer
Albert Lee¹, Yi Xu²
albertlee@eduhk.hk, yi.xu@ucl.ac.uk

Analysis-by-synthesis is a useful way to evaluate theoretical models of speech prosody and to understand their workings and assumptions, which are usually not clear to the novice researcher. This approach complements traditional acoustic analysis and can offer additional insights to the production data in hand. The present study extends previous f₀ modelling work (Lee & Xu, 2015; Lee, Xu, & Prom-on, 2014) based on an articulatory model of speech prosody (Xu, 2005) and explores predictive synthesis based on the Autosegmental-Metrical Theory (e.g. Pierrehumbert & Beckman, 1988, AM henceforth).

The corpus consists of 2,640 utterances produced by eight native speakers of Japanese (Lee, Prom-on, & Xu, 2017). These utterances are single words framed in an unaccented carrier sentence, and contrast in word length and pitch accent conditions. We used the f₀ synthesizer AMtrainer (ver. 3.8.1) to interpolate f₀ contours between AM tone targets through (i) local resynthesis and (ii) the Jackknife procedure. The former refers to re-interpolating an f₀ contour between AM tone targets on an individual utterance, whereas the latter was done by averaging the peak delay times of utterances by all speakers other than the one being evaluated.

Preliminary (N = 1,979) local resynthesis accuracy results were comparable to a previous study (Lee et al., 2014), RMSE = 1.026, Pearson’s r = .920, with unaccented words achieving lower accuracy than accented ones. That of predictive synthesis using the Jackknife procedure was almost as high, RMSE = 1.180, Pearson’s r = .906. One-tailed paired t-tests revealed that the differences between the two synthesis approaches in both RMSE and r were significant, respectively t(1977) = -15.559, p < .001, and t(1977) = 6.442, p < .001, albeit small.

These findings demonstrate that (i) AM is capable of yielding very good predictive accuracy synthesis and that (ii) the temporal alignment of AM tone targets is highly stable even across speakers in Japanese. Our work paves the way for future work comparing multiple theoretical models of speech prosody based on the same evaluation metrics and speech corpora.

References