Post-collection synchronization of ultrasound and audio

Sarah Bakst [1], *Susan Lin*[2]

[1] University of Wisconsin–Madison, USA
[2] University of California, Berkeley, USA

sbakst@wisc.edu, susanlin@berkeley.edu

Background: Speech research using lingual ultrasound often requires both the ultrasound images as well as audio from the corresponding audio speech signals, but synchronization of these signals is not always available. We propose that periods of matching rates of change in the two signals could be used to align articulatory and acoustic signals where synchronization is impossible or would otherwise benefit from verification.

Methods: We analyzed ultrasound (100 fps) and audio recordings of 8 English speakers reading real CV(C) words (59 stimuli, 8-9 repetitions) in the frame "I'm a ___ ." Audio and ultrasound streams were synchronized at the time of capture via hardware synchronization signal. Each frame of ultrasound data was represented as a matrix of pixel brightness values, and articulatory change was calculated as the mean of the absolute value of their difference matrices over time. Acoustic change was similarly calculated as the mean of the absolute value of the difference between Mel frequency cepstral coefficients representations of the audio recording. Segmental boundaries were determined using the Penn Forced Aligner (Yuan and Liberman 2008) on the audio alone. Figure 1 shows relative articulatory and acoustic change for the utterance "I'm a Lee." We then calculated the degree of correlation between acoustic and articulatory change for windows of 150, 180, 210, and 240ms. Figure 2 shows these r- and p-values for the same utterance. Then we deliberately offset the signals by ±150, ±100, ±50, ±30, ±20, and ±10 ms to verify that the known synchronization results in the best correlations. Figure 3 shows average r-values for all data from all subjects.

Results and discussion: Preliminary results suggest that this method is generally successful: offsets closer to zero produce the best correlation between signals. Shorter window lengths improve alignment, as do restricting analysis of correlations to periods of detected speech.

Conclusions: This method shows promise for future use in aligning signals that lack a synchronization pulse. However, for small offsets, the differences in correlation coefficient are not very large; future work will investigate optimal lengths of speech for determining alignment. Ongoing analysis to perfect this tool includes determining whether duration of correlation (number of windows with high r-values) or overall degree of correlation (median r-values) leads to the most accurate alignments.

References

Jadoul, Y. et al. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71, 1-15.

Stevens, K. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. *Human Communication: a unified view*.

Yuan, J. & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics* '08.

Keywords: Methodological research, Speech production, Articulatory-acoustic relations

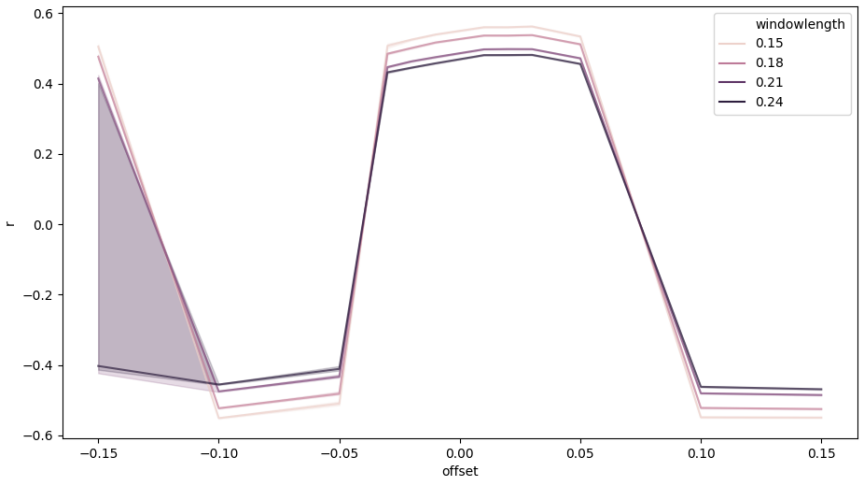

Figure 3: *Average correlation between articulatory and acoustic change for different windows (color) and offsets between signals (x-axis). Only correlations that are significant (p<0.05) are plotted. Highest correlations occur near 0-offset. Offset is in seconds.*

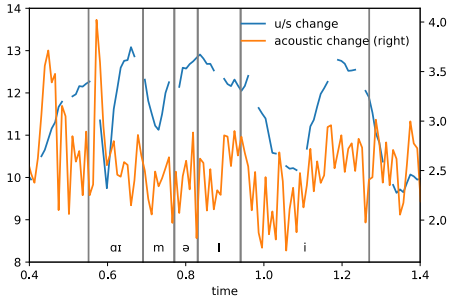

Figure 1: *Change in acoustics (orange) and articulation (blue) for the utterance "I'm a Lee." Time is in seconds.*

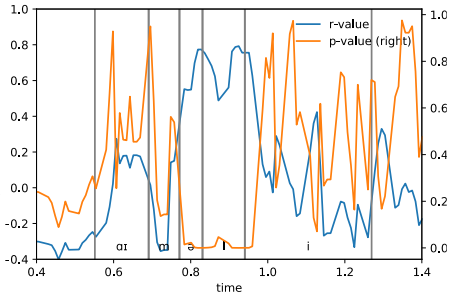

Figure 2: *R- (blue) and p-values (orange) of the correlation between the articulatory and acoustic signals shown in Figure 1.*