Duration of connected speech needed to accurately estimate the Articulatory-Acoustic Vowel Space of a reading passage

Jason Whitfield[1], Ph.D., Anna C. Gravelin[1], M.S., Zoe Kriegel[1], B.S., & Daryush Mehta[2], Ph.D.
[1]Communication Sciences and Disorders, Bowling Green State University;
[2]Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital; [3]Department of Surgery; Harvard Medical School

Motor Speech Lab
Speech Production & Motor Learning

Introduction

Vowel Space metrics calculated from formant frequencies have been used as acoustic measures of vowel articulation in a variety of applications, including characterizing the effect of dysarthria on articulation, characterizing dialectical variation, examining changes in vowel articulation associated with development, and characterizing the effect of style-related changes (e.g., clear speech) on speech acoustics (e.g., Jacewicz, et al., 2007; Lam et al., 2012; Vorperian & Kent, 2007). Recent work has examined utterance-level vowel space metrics that are based on continuously sampled formant trajectories extracted from connected speech. Unlike other vowel space metrics, the Articulatory-Acoustic Vowel Space (AAVS; Whitfield & Goberman, 2014, 2017) is calculated from the generalized variance of continuously sampled formant (F1-F2) traces.

Given a sufficiently long speech sample, the AAVS should stabilize, though the duration required for the measure to converge remains unknown. The current investigation aimed to determine the amount of formant data needed for the AAVS to stabilize.

Method

Formant traces were extracted from readings of the Caterpillar passage (Patel et al., 2013) produced by 16 speakers using habitual speech (Mean age=64.93 yrs; Range=48 to 81 yrs).

Time series (20 ms frames, 10 ms intervals) of the first three formant frequencies were extracted using a Kalman-based autoregressive approach consisting of three stages: pre-processing, intra-frame observations, and inter-frame tracking (Mehta et al., 2012). PRAAT-based voice activity detection was used to identify voiced and voiceless portions of each signal to ensure that formant trajectories from the voiced portion were analyzed. For each voiced interval, any local outliers in the formant trace were removed using a median absolute deviation moving average function in MATLAB. The formant trace was then low-pass filtered at 10 Hz. As a final step, bivariate outliers in the filtered formant trajectory were identified by calculating the Mahalanobis distance for each F1-F2 pair. F1-F2 pairs that were greater than 2 standard deviations from the centroid were removed from the trace.

The passage AAVS was calculated as the square root of the generalized variance of the F1-F2 data (Whitfield & Goberman, 2014; 2017). The absolute percent difference (error) between cumulative AAVS estimates and the passage AAVS was calculated by iteratively adding F1-F2 pairs from successive voiced frames. Power functions were fit to the error plots using a least absolute residuals method to determine the frame at which the upper bound of the function fell within 5% of the passage AAVS value.

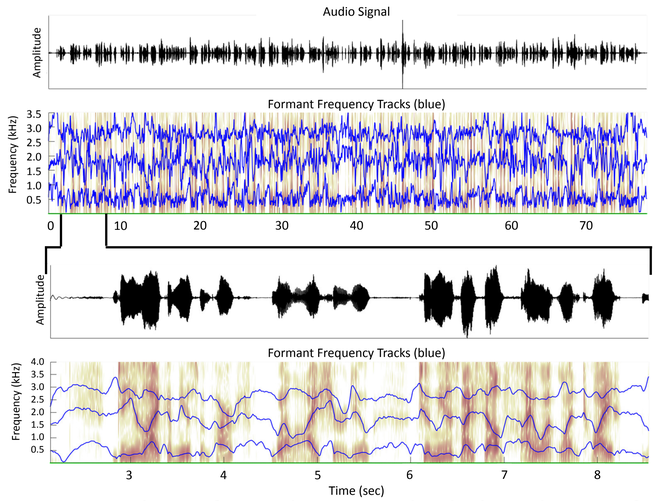

Figure 1. Example waveform and spectrographic display with KARMA formant frequency trace (blue) for the full Caterpillar passage (top) and an expanded segment of the passage (bottom). Note: Only formant frequency traces from the voiced frames were used for the AAVS calculations.

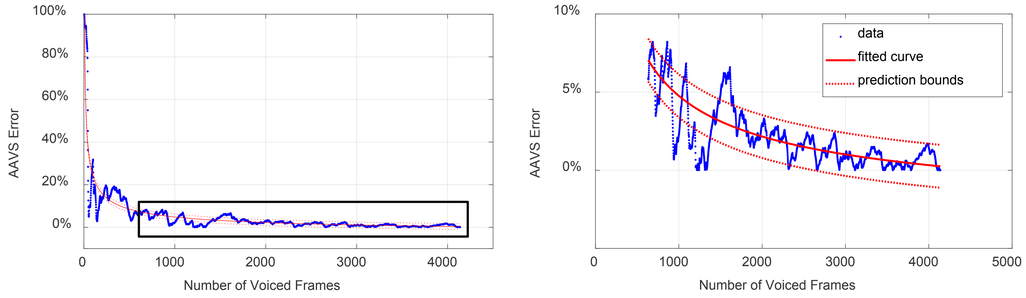

Figure 2. Example power function from one participant displaying that the error between the AAVS estimates and passage value decreases when F1-F2 pairs are added to the measure, showing convergence of the AAVS estimates.

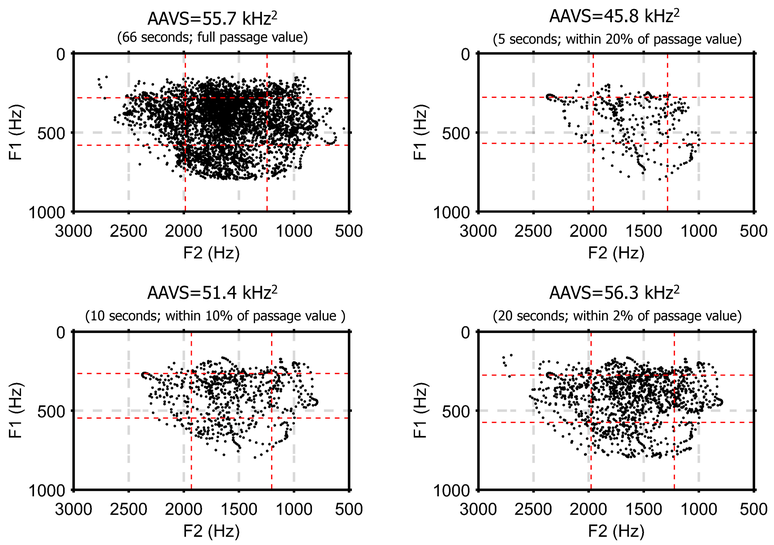

Figure 5. Visual representations of the Articulatory-Acoustic Vowel Space (AAVS) calculated for the passage (top left pane), the first 5 seconds of the sample (top right pane), the first 10 seconds of the sample (bottom left pane), and the first 20 seconds of the sample (bottom right pane). Dashed read lines indicate +/- 1 standard deviation from the mean F1 and F2.

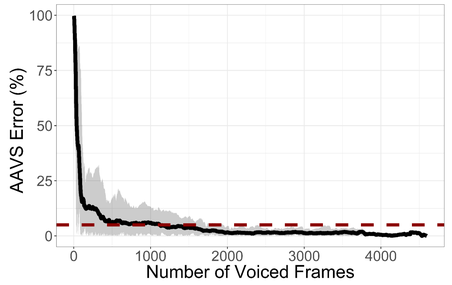

Figure 3. Group-level error estimates representing the relative difference in the AAVS estimates from the passage value as a function of number of samples included in the measure. The red line represents the 5% error level.

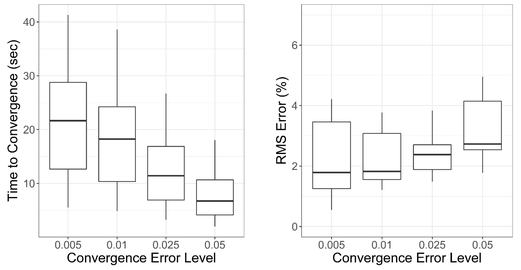

Figure 4. Box and whisker plots for the speaking duration needed for AAVS convergence (left) and percent root-mean mean square error of the AAVS after convergence (right) for four convergence levels based on the error (i.e., relative difference) between the AAVS estimates and the passage AAVS value.

Results & Discussion

Across all speakers, the AAVS converged within 1065 voiced frames (i.e., 10.65 seconds of voiced speech). In terms of absolute speaking duration, all samples converged within 18 seconds (Mean=8.26; SD=5.04). These data suggest that 15 to 20 seconds of connected speech is sufficient to provide a reasonable estimate of working vowel space using the AAVS. Based on the current data, examining changes in vowel space across a speaking task seems feasible, given that the sample is sufficiently long (e.g., 60 seconds) and the phonetic content is relatively balanced.

References

Jacewicz, E., Fox, R. A., & Salmons, J. (2007, August). Vowel space areas across dialects and gender. In International Congress of Phonetic Sciences (Vol. 16, pp. 1465-1468).
Lam, J., Tjaden, K., & Wilding, G. (2012). Acoustics of clear speech: Effect of instruction. Journal of Speech, Language, and Hearing Research, 55(6), 1807-1821.
Mehta, D. D., Rudoy, D., & Wolfe, P. J. (2012). Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking. The Journal of the Acoustical Society of America, 132(3), 1732-1746.
Patel, R., Connaghan, K., Franco, D., Edsall, E., Forgit, D., Olsen, L., Ramage, L., Tyler, E., Russell, S., (2013). "The Caterpillar": A novel reading passage for assessment of motor speech disorders. *American Journal of Speech-Language Pathology, 22*(1), 1-9.
Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. Journal of Speech, Language, and Hearing Research, 50(6), 1510-1545.
Whitfield, J. A., & Goberman, A. M. (2014). Articulatory–acoustic vowel space: Application to clear speech in individuals with Parkinson's disease. Journal of communication disorders, 51, 19-28.
Whitfield, J. A., & Goberman, A. M. (2017). Articulatory-acoustic vowel space: Associations between acoustic and perceptual measures of clear speech. International journal of speech-language pathology, 19(2), 184-194.

175th Meeting of the Acoustical Society of America, 2018, Minneapolis, MN
jawhitf@bgsu.edu