Examining speaker variability using low-dimensional and high-dimensional phonetic representations

Speech events are unique; speakers do not produce the same sound in exactly the same way twice. They vary their speech depending on a whole range of factors – speaker internal (emotion, health, etc.) and speaker external (interlocutor, topic, environment, etc.). Intra-speaker variation is significant because it is a leading cause of incorrect speaker identification (Zhang et al, 2006), shows socially-meaningful patterning (Podesva, 2007), and can represent potential cues to the origin and spread of sound change (Mielke et al. 2019). Holistically tracking speaker variability is, however, very challenging. For example, different linguistic features may show different degrees of variability and different measurements may also produce different conclusions (Rhodes 2012). This study aims to address these issues by examining the accuracy of speaker classification across multiple samples per speaker, focussing on comparing different phonetic representations.

A speaker classification experiment was conducted on 20 male speakers aged 18-24, from two UK dialects (Manchester and Newcastle; Haddican & Foulkes 2017). Three 30-second spontaneous speech samples were extracted for each speaker at three different time points from within the same recording, which allows us to examine the robustness of different speaker modelling methods in light of within-speaker variation. Specifically, we compare vowel formants (a low-dimensional representation) and 13 MFCCs (a high-dimensional representation) for each speaker in order to observe how the multiple samples from each speaker cluster together across these representations. Clustering was performed using Gaussian Mixture Models (GMMs) and agglomerative cluster analyses, while the success of each model was assessed in terms of how accurately it classified speech samples from the same speaker.

The results show that the extent of intraspeaker variation is sufficient to inhibit accurate speaker classification. MFCCs performed better than formant measurements in identifying contemporaneous samples within speakers, although the effect of formants vs MFCCs was also variable between speakers. This points towards differential weighting of information between speakers in determining speaker individuality. These results are discussed in terms of the extent of speaker variability and the need for greater interpretability in high-dimensional feature sets. I further outline some remaining challenges in the study of intra-speaker variation and its relevance to applied phonetics.

Haddican, William and Foulkes, Paul (2017). *A comparative study of language change in Northern Englishes.* [Data Collection]. Colchester, Essex: ESRC.

Mielke, J., Thomas, E. R., Fruehwald, J., McAuliffe, M., Sonderegger, M., Stuart-Smith, J., & Dodsworth, R. (2019). Age vectors vs. axes of intraspeaker variation in vowel formants measured automatically from several English speech corpora.

Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of sociolinguistics*, *11*(4), 478-504

Rhodes, R. W. (2012). *Assessing the strength of non-contemporaneous forensic speech evidence* (Doctoral dissertation, University of York).

Ross, S., Earnshaw, K., & Gold, E. (2019, August). A Cautionary Tale for Phonetic Analysis: The Variability of Speech Between and Within Recording Sessions. In *19th International Congress of the Phonetic Sciences* (pp. 3090-3094).

Zhang, C., van de Weijer, J., & Cui, J. (2006). Intra-and inter-speaker variations of formant pattern for lateral syllables in Standard Chinese. *Forensic science international*, *158*(2-3), 117-124.