Between-speaker variability in dynamic formant characteristics in spontaneous speech

Lei He[1] and Willemijn Heeren[2]
[1] *Department of Computational Linguistics, University of Zürich*
[2] *Leiden University Centre for Linguistics, Leiden University*
lei.he@uzh.ch; w.f.l.heeren@hum.leidenuniv.nl

Introduction

The temporal characteristics of speech articulation have received relatively little attention in forensic phonetics, because directly characterizing speaker-specific articulatory movements is almost impossible; kinematic data of articulators are absent from case materials. However, forensic speech scientists may instead focus on acoustic properties in the speech signal that are – although not entirely – modulated by the articulatory movements. For example, Dellwo and colleagues measured speech rhythm in terms of the durational variability of various phonetic intervals (e.g., Dellwo et al. 2015, Leemann et al. 2014) or syllabic intensity variability (e.g., He and Dellwo 2014, 2016); McDougall (2006) approached formant trajectories using least-squares polynomial approximations; and He and Dellwo (2017) measured the dynamic characteristics of intensity contours. Their study found that measures based on the speeds of intensity decreases (i.e., negative intensity dynamics) explained approximately 70% of between-speaker variability, pointing to a possibility that the mouth-closing gestures may contain more speaker-specific information.

More recently, He et al. (2019) combined the ideas of both McDougall (2006) and He and Dellwo (2017) and measured the dynamic characteristics of the first formant (F1). They found that the speeds of F1 decreases (reflecting mouth closing movements) contained more speaker-specific information than speeds of F1 increases (reflecting mouth opening movements). Moreover, an advantage of using F1 over intensity is that F1 measures are less affected by varying distances to the microphone. This is particularly relevant in forensic scenarios; voice experts typically have no information about the mouth-to-transducer distance, and distance may vary, in an unknown way, in the course of a recording. Moreover, the result that measures of negative F1 dynamics explained more between-speaker variability than measures of positive F1 dynamics is highly congruent to He and Dellwo (2017) using intensity dynamics.

However, He et al. (2019) only focused on Zürich German read speech in laboratory settings. To evaluate the practical value of this method for forensic practices, the current research aimed to test whether the same results will be obtained using spontaneous speech, in different languages. Thus, we aimed to investigate the generalizability of the findings from He et al. (2019) to scenarios much closer to the ones found in forensic speaker comparisons.

Method

Corpora and speakers

Vocalic nuclei were manually annotated in Praat (Boersma and Weenink, 2017) in data from three corpora, in different languages. This was done using phonetic transcripts created through forced alignment of available orthographic transcripts:

– For English, telephone conversations from 14 speakers were annotated (DyVis corpus [Nolan 2011], task 2). Per speaker, between 26 and 40 sentences were included (M = 33);

– For Dutch, spontaneous face-to-face conversations from 16 gender-balanced speakers were included (Spoken Dutch Corpus http://lands.let.ru.nl/cgn/ehome.htm). Per speaker, between 25 and 43 sentences were included (M = 34);

– For Zürich German, the TEVOID (Dellwo et al. 2015) corpus was used, containing16 gender-balanced speakers. Per speaker, 16 spontaneous sentences were extracted from an interview with an experimenter.

Acoustic and statistical analysis

The trajectories of F1 of each syllable nucleus were extracted using Praat (Boersma & Weenink, 2017), and the F1 dynamics (F1[+] and F1[–]) were calculated following the procedure described in He et al. (2019). The distributional characteristics of F1[+] and F1[–] in each sentence were calculated in terms of the mean (mean_F1[+] and mean_F1[–]), the standard deviation (stdev_F1[+] and stdev_F1[–]) and pairwise variability index (pvi_F1[+] and pvi_F1[–]). Multinomial logistic regressions were used to test the amount of between-speaker variability each of these measures can explain. This procedure was repeated for each of the languages.

Data processing and analysis are currently under way. We will present and discuss the results at the conference.

Acknowledgements

This work is being supported by an IAFPA research grant and an NWO VIDI grant (276-75-010).

References

Boersma, P; Weenink, D (2017) “Praat: doing phonetics by computer,” Version 6.0.28, downloaded from http://www.fon.hum.uva.nl/praat/.

Dellwo, V; Leemann, A; Kolly, M-J (2015) “Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors,” *Journal of the Acoustical Society of America* 137: 1513–1528.

He, L; Dellwo, V (2014) “Speaker idiosyncratic variability of intensity across syllables,” in *Interspeech 2014* , Singapore, pp. 233–237.

He, L; Dellwo, V (2016) “The role of syllable intensity in between-speaker rhythmic variability,” *International Journal of Speech, Language and the Law* 23: 243–273.

He, L; Dellwo, V (2017) “Between-speaker variability in temporal organizations of intensity contours,” *Journal of the Acoustical Society of America* 141: EL488–EL494.

He, L; Zhang, Y; Dellwo, V (2019) “Between-speaker variability and temporal organization of the first formant” *Journal of the Acoustical Society of America* 145: EL209–EL214.

Leemann, A; Kolly, M-J; Dellwo, V (2014) “Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison,” *Forensic Science International* 238: 59–67.

McDougall, K (2006) “Dynamic features of speech and the characterisation of speakers: Towards a new approach using formant frequencies,” *International Journal of Speech, Language and the Law* 13: 89–126.

Nolan, F (2011) Dynamic Variability in Speech: a Forensic Phonetic Study of British English, 2006–2007 [data collection], UK Data Service.