Between-speaker variability in intensity slopes: The case of Thai

Yu Zhang[1], Lei He[1], Karnthida Kerdpol[2], Volker Dellwo[1]

[1]Phonetics & Speech Sciences Group, Institute of Computational Linguistics, University of Zurich, Andreasstrasse 15, CH–8050, Zurich, Switzerland
[2]Department of Linguistics, Naresuan University, Phitsanulok 65000, Thailand

Introduction

The major processes in speech production - glottal vibrations and articulatory movements - all contain speaker idiosyncratic information; such speaker idiosyncrasies leave traces in the speech signal and are thus measurable acoustically (see Dellwo et al., 2019 for a general review). Time-invariant properties of speech such as average fundamental frequencies (f0) and formant center frequencies have been extensively charted in speaker individuality research, and thus are generally accepted parameters of speaker identification decisions in forensic voice comparison and automatic speaker recognition (Singh and Murry 1978; Jessen 1997; Adami, Mihaescu, Reynolds and Godfrey 2003; Nolan and Grigoras 2005; Zhang, van de Weijer and Cui 2006; Lindh and Eriksson 2007; Morrison 2009; Leemann, Mixdorf, O'Reilly, Kolly and Dellwo 2014). However, speakers must evidently present some particular temporal features in the speech signals due to individual ways of moving their articulatory apparatus. He and his colleagues (2017, 2019) used positive and negative slopes to indicate both the speeds of intensity change and F1 change in a sentence and demonstrated that measures of negative slopes could explain more between-speaker variability in Zurich German (≈70%). Since intensity slopes and F1 slopes are both modulated by the opening-closing gestures of the mouth, these two acoustic measures are thus good estimates of mouth articulatory movements. Congruency in the findings indicates that the mouth-closing gestures during speech articulation may encode more speaker-specific information. Nevertheless, the previous findings (He and Dellwo, 2017; He et al., 2019) were obtained solely from native speakers of Zurich German. Investigating whether similar findings can be replicated from speakers of other languages is of great importance, especially when we wish to link the theoretical implications to forensic caseworks involving speakers of many languages. This paper thus followed the method with an improved statistical analysis and looked at how speaker differences are manifested in the temporal organizations of signal intensity curve using a Thai speech corpus.

Method

Thirteen native speakers of standard Thai (all female, aged between 20 to 22) were recorded reading the same set of 355 sentences (unidirectional microphone, sound-treated booth at Naresuan University, Phitsanulok/Thailand; 44.1 kHz, 16-bit). Following the same procedure in He and Dellwo (2017), positive and negative slopes (V[+] and V[–]) were calculated and then mean, variation coefficient (varco) and pairwise variability index (pvi) were used to characterize the distributions of positive slopes and negative slopes in a sentence. Z-score normalizations were performed for each particular measure to control sentence effect. The variance inflation factor (VIF) was computed for all six intensity slopes measures for diagnosing collinearity. The multinomial logistic regression (MLR) was fitted to quantify the amount of between-speaker variability explained by each of the intensity slope measures.

Results and discussion

The MLR results show that collectively measures of negative slopes explained 65.60% between-speaker variability (significantly higher than 50%, $\chi^2_{(1)} = 9.74$, $p < .01$). To directly compare the results with what He and Dellwo (2017) reported, we reanalyzed the Zurich German data and the MLR results showed that measures of negative slopes explained

68.44% (significantly higher than 50%, $\chi^2_{(1)} = 13.60$, $p < .001$) between-speaker variability in Zurich German, indicating that the results from the two languages were very similar.

The suprasegmental intensity or sonority fluctuations are one of the acoustic outcomes of mouth opening-closing movements, which organize speech into syllable-sized units constituting the rhythmic frames; this process has been argued to have an evolutionary advantage in speech comprehension (e.g., Chandrasekaran et al., 2009; MacNeilage, 1998; Morrill et al., 2012; Strauss and Schwartz, 2017). It is thus likely that the role of mouth opening-closing cycles is universal across languages, and the way speaker-specificity is encoded in the dynamic process and its acoustic outcomes are also similar in different languages. Needless to say, more languages should be investigated to testify our interpretation.

The findings have implications for research and applications where identity information in speech matters, such as forensic voice comparison (FVC) and automatic speaker recognition (ASR).

References

Adami, A., Mihaescu, R., Reynolds, R.D. and Godfrey, J.J. (2003). Modeling prosodic dynamics for speaker recognition, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, 788-91.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. and Ghazanfar, A. (2009). The natural statistics of audiovisual speech. PLoS Computational Biology 5, e1000436.

Dellwo, V., French, P., and He, L. (2019). "Voice biometrics for speaker recognition applications," in The Oxford Handbook of Voice Perception, edited by S. Frühholz and P. Belin (Oxford University Press, Oxford, UK), pp. 777--795.

He, L. and Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. The Journal of the Acoustical Society of America 141, EL488--494.

He, L., Zhang, Y., and Dellwo, V. (2019). "Between-speaker variability and temporal organization of the first formant," J. Acoust. Soc. Am. 145, EL209--EL214.

Jessen, M. (1997). Speaker-specific information in voice quality parameters. Forensic Linguistics 4, 84-103.

Leemann, A., Mixdorff, H., O'Reilly, M., Kolly, M-J. and Dellwo, V. (2014). Speaker-individuality in Fujisaki model f0 features: implications for forensic voice comparison. International Journal of Speech, Language and the Law 21(2), 343-370.

Lindh, J. and Eriksson, A. (2007). Robustness of long time measures of fundamental frequency. Proceedings of INTERSPEECH 2007, 2025-2028.

MacNeilage, P. F. (1998). "The frame/content theory of evolution of speech production," Behav. Brain Sci. 21, 499--546.

Morrill, R. J., Paukner, A., Ferrari, P. F., and Ghazanfar, A. A. (2012) "Monkey lipsmacking develops like the human speech rhythm," Develop. Sci. 15, 557--568.

Morrison, G. S. (2009). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. The Journal of the Acoustical Society of America 125, 2387-2397.

Nolan, F. and Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. International Journal of Speech, Language and the Law 12, 385-411.

Singh, S. and Murry, T. (1978). Multidimensional classification of normal voice qualities. The Journal of Acoustical Society of America 64, 81-87.Strauss, A., and Schwartz, J.-L. (2017). "The syllable in the light of motor skills and neural oscillations," Lang. Cogn. Neurosci. 32, 562--569.

Zhang, C., van de Weijer, J. and 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.