Using Phonetic Theory to Improve Automatic Speaker Recognition

Elliot Holmes (University of York, ejh621@york.ac.uk)

Applications of Automatic Speaker Recognition are now widespread, being used in forensic voice comparison cases around the world whilst global commercial organisations (such as banks) use it for customer verification purposes. Within the field, there is a growing focus on the role that phonetic theory can play in complementing Automatic Speaker Recognition, specifically for understanding and improving upon current 'black box' approaches which developers cannot interpret. Such approaches have become popular in many other fields due to their increased speed and performance over interpretable systems and remain popular today; however, they are also proving increasingly problematic. For example, Rudin (2019) observed a 'black box' system used in hospitals to assess patient risk that, when it failed, put lives at stake. As the system was uninterpretable, the problem could not be identified nor rectified; however, once replaced with an interpretable system (which performed as well as the 'black box' system originally did), hospital staff are now able to understand any errors produced and rectify them.

Automatic Speaker Recognition does not need to take a similar risk; interpretable, phonetic features can be used to recognise speakers. Zhu et al. (2009) found that pitch features can discriminate between speakers; Long et al. (2011) found that harmonics-to-noise ratio can, and Ali et al. (2006) found that formants and bandwidths can. Studies have also shown that integrating interpretable phonetic features into current 'black box' systems, in particular those that measure laryngeal voice quality, can improve performance (Hughes et al., 2019).

The paper first presents a new methodology for testing phonetic approaches to improving Automatic Speaker Recognition. Current 'black box' approaches do not tailor, analyse, or specify any phonetic features or approaches before conducting Automatic Speaker Recognition tasks. The proposed methodology, however, is novel in that it specifies and extracts measurements of 35 phonetic features and measures them across phonetic units (phonemes) taken from every speech signal in two corpora. Then, comparisons are made between the two corpora, the contents of which can be tailored to the investigation at hand: this might be a comparison of two individual speakers, two gender groups, two age groups, or two accent groups. These two data sets are compared on a feature-by-feature, phoneme-by-phoneme basis using statistical modelling, specifically Linear Mixed Models (LMMs), to identify which linguistic features in each phoneme distinguish the two data sets. LMMs are more robust, as they allow for random effects structures to be specified. The aim is, in essence, feature selection, whereby we assess and rank features according to how often they are able to separate pairs of speakers in our dataset.

Then, this paper will utilise this methodology for a short study to demonstrate its usefulness: it will distinguish 100 speakers using the 35 features to compare tokens of their production of /a/. Specifically, every speaker was individually compared to the other 99 speakers using LMMs to identify which features distinguish their production of /a/. Nolan et al.'s (2009) DyViS Corpus has been selected for this study because there are 100 participants, all of which are comparable along sociolinguistic axes (gender, age, accent) and recorded with the same microphone. All tokens of the /a/ phoneme are text-dependent, having been taken from their production of the same passage of text.

This study presents findings that move towards a linguistically-informed improvement to Automatic Speaker Recognition: the methodology is novel and interpretable and its results identify phonetic

features in specific phonemes that are salient in distinguishing and characterising speakers from one another. Importantly, this methodology can now be replicated using bigger databases to expand upon the present example sample or even to identify distinguishing features between groups of speakers based on age, gender, and accent. Furthermore, it could also be used on shorter utterances or on text-independent speech to identify which phonetic features are salient to contexts more attuned to real-world uses of Automatic Speaker Recognition systems.

Reference List

Ali, A., Bhatti, S., and Mian, M. S. (2006, April, 22-23). *Formants Based Analysis for Speech Recognition*. [Paper]. IEEE International Conference on Engineering of Intelligent Systems (ICEIS), Islamabad, Pakistan. https://ieeexplore.ieee.org/ .

Hughes, V., Cardoso, A., Harrison, P., Foulkes, P., French, P., and Gully, A. J. (2019). *Forensic voice comparison using long-term acoustic measures of voice quality.* [Paper]. International Conference of Phonetic Sciences, Melbourne, Australia. https://vincehughes.files.wordpress.com/.

Long, Y., Yan, Z., Soong, F. K., Dai, L., Guo, W. (2011, May, 22-27). *Speaker characterization using spectral subband energy ratio based on Harmonic plus Noise Model.* [Paper]. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic. https://ieeexplore.ieee.org/.

Nolan, F., McDougall, K., de Jong, G., and Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *Forensic Linguistics,* 16(1). https://www.researchgate.net/.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence,* 1, 206-215. https://doi.org/10.1038/s42256-019-0048-x

Zhu, J., Sun, S., Liu, X., and Lei, B. (2009). *Pitch in Speaker Recognition.* [Paper]. Proceedings of the 2009 Ninth International Conference on Hybrid Intelligent Systems, Massachusetts, U.S.A. https://dl.acm.org/.