Seeing the trees in the forest: Diagnosing individual performance in likelihood ratio based forensic voice comparison

Justin J. H. Lo (University of York, UK)
jl2355@york.ac.uk

Introduction

In numerical likelihood ratio (LR) based forensic voice comparison (FVC), system testing is predominantly evaluated on the global level. System validity is assessed by means of a single metric score representing the proportion or size of errors, the most commonly used of which are the equal error rate (EER) and log likelihood ratio cost function (C_{llr}). Other graphical means of illustration, such as ROC curves, DET curves and Tippett plots, are also regularly used to provide more information about the overall performance of the system.

As much as such measures and graphs can provide an indication of global system performance, their diagnostic value can be limited. Within any given system, variation of performance between individual speakers may arise due to physiological, behavioural and technical reasons. Analysis of performance on the level of individual speakers thus offers the potential to gain insights into the nature of errors and, as such, for work towards improving system design in a targeted manner. In addition, an examination of individual performance in relation to the acoustic-phonetic data used to generate output LRs can further our understanding of the relationship between input and output in LR based FVC.

As analysis of individual performance remains rarely performed in the context of FVC, the present study seeks to demonstrate its utility in LR based FVC and further explores the connection between individual performance as derived from LRs and the underlying speech data. As a case study, this study makes use of long-term formant distributions (LTFDs), a set of features argued to be able to capture both anatomical variation of the vocal tract and idiosyncratic articulatory habits of speakers.

Methods

The present study is part of an ongoing project that draws its data from the Voice ID Database (RCMP 2010–2016). High-quality microphone recordings from 60 adult male bilingual speakers of Canadian English and French were analysed, but the current exploratory analysis is limited to the English materials, consisting of phonetically balanced read sentences and passages. All recordings were automatically segmented using the Montreal Forced Aligner (McAuliffe et al. 2017), followed by manual checks and corrections where necessary. Formant estimates for the first four formants were automatically extracted in Praat at 10ms intervals from all vowels and glides, with the formant tracker set to search for 6 formants up to a maximum formant frequency of 5500 Hz in 25ms frames.

To evaluate the performance of LTFDs as speaker discriminants, speakers were randomly partitioned into three equally sized of test, training and reference speakers. LTFDs from all four formants combined were modelled and compared using GMM–UBMs (Reynolds et al. 2000) to generate scores, which were then calibrated by logistic regression to obtain $\log_{10}$LRs (LLRs). LTFDs from each formant were likewise tested separately to facilitate one-to-one comparison with the acoustic data. C_{llr} and EER were calculated to assess the overall performance of each system. The sampling and testing procedure was replicated 100 times, in order to minimise the effects of random speaker sampling on the comparison of LLRs and metrics of validity.

For analysis on the individual level, this study makes use of the notion of the "biometric menagerie" (Doddington et al. 1998), where speakers are classified into user groups, or animals, based on their performance. Zooplots were constructed by plotting a speaker's average performance in different-speaker comparisons against their performance in same-speaker comparisons. Average performance of any speaker in this study is defined as the arithmetic mean of LLRs from all same- or different-speaker comparisons (SS-LLR; DS-LLR) involving that speaker. Following the definitions in Dunstone and Yager (2009), speakers whose performance ranked within the top or bottom quartile of all speakers, in both same-speaker and different-speaker comparisons, were identified and classified as doves, worms, phantoms and chameleons, and their LTFD data were further analysed with respect to other speakers in the group.

Results and discussion

In terms of global system performance, all systems using LTFDs from single formants reported similar mean C_{llr} (0.62-0.69) and EER (19-22%), suggesting that overall each LTFD performed at a similar level. The system combining LTF1-4 performed considerably better, as evidenced by the lower mean C_{llr} (0.31) and EER (7%).

Zooplots show that, for each formant, all speakers reported a negative mean DS-LLR, indicating that, on average, they were capable of being distinguished from other speakers. While the majority of speakers had a positive mean SS-LLR, some speakers reported negative mean SS-LLRs, indicating same-speaker comparisons as a source of errors.

Zoo analysis further shows that, among the set of 60 speakers, different subsets of speakers were identified as doves and worms for each individual formant, thus providing corroborating evidence on the individual level that each LTFD captures some complementary speaker-specific information. Comparison with the underlying acoustic data shows clear separation between doves and worms, especially in LTF3 and LTF4. The distributions of doves typically exhibited peaks at more extreme frequencies, while those of worms had similar shapes to the overall distribution of the group.

The findings above demonstrate the diagnostic value that individual-level analysis can add to system evaluation in LR based FVC by providing a more fine-grained picture of performance. Further analysis of the acoustic-phonetic data illustrates how individual distributions of acoustic-phonetic data can be reflected in exceptional speaker-discriminatory performance. This study thus supports an approach where LR based FVC is concerned with not only global measures of validity and reliability, but also the performance of individual speakers and the factors behind their performance.

References

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 Speaker Recognition Evaluation.

Dunstone, T. & Yager, N. (2009). *Biometric system and data analysis: Design, evaluation, and data mining*. New York: Springer.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). *Montreal Forced Aligner* [computer programme]. Version 1.0.0, retrieved 30 November 2017 from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, 12(2), 143–173.

Reynolds, D., Quatieri, T., & Dunn, R. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3), 19–41.

Royal Canadian Mounted Police [RCMP]. (2010–2016). Voice ID Database [unpublished audio corpus]. Collected at University of Ottawa.