Sensitivity of x-vectors and automatic speaker recognition scores to vocal variation

Jessica Wormald[1], Paul Foulkes[1], Philip Harrison[1], Vincent Hughes[1], Finnian Kelly[2], David van der Vloed[3], Poppy Welch[1], Chenzi Xu[1]

[1]Department of Language and Linguistic Science, University of York, UK
[2]Oxford Wave Research, Oxford, UK
[3]Netherlands Forensic Institute, The Hague, The Netherlands
{jessica.wormald|paul.foulkes|philip.harrison|vincent.hughes|poppy.welch|chenzi.xu}@york.ac.uk,
finnian@oxfordwaveresearch.com, d.van.der.vloed@nfi.nl

ABSTRACT

Automatic speaker recognition (ASR) systems rely on a complex processing chain in order to compare speech signals and produce likelihood ratios. The complexity of this chain, and of the speech signals themselves, mean that there is still limited understanding about what makes a certain voice easy or difficult for a system to recognise. This gap in understanding is holding back the use of ASR in forensic casework.

This study considers two specific parts of the ASR chain: x-vectors (speaker models) and within-speaker comparison scores. Using heavily-controlled data from two phoneticians, we demonstrate that variability in vocal setting results in phonetically-predictable shifts in x-vectors and scores. Shifts in supralaryngeal voice quality produce the biggest deviations from modal voice. The results provide a basis for exploring how properties of the voice affect ASR performance, which in turn can contribute to helping courts and practitioners take advantage of ASR systems in forensic casework.

Keywords: speaker recognition; forensic phonetics; vocal setting

1. INTRODUCTION

1.1 The context

Automatic speaker recognition (ASR) systems are increasingly used around the world in forensic speaker comparison cases [1,5]. State-of-the-art systems utilise deep neural networks (DNNs) to convert acoustic features (typically mel frequency cepstral coefficients, MFCCs) to a compact, fixed-length speaker representation, known as an x-vector [8]. Such systems perform very well, even in forensically realistic conditions, when optimised using case-specific data [4,6].

While much research in ASR has focused on tackling technical challenges arising from different recording types (e.g. channel effects such as telephone transmission, background noise), still very little is known about why certain voices are easy or difficult for systems to recognise. This, in part, explains why some courts are still cautious about ASR as a form of expert evidence [7]. A particular concern is being able to explain what information about the voice is being captured by complex models like DNNs. This study is the first step in a wider project which will provide a more comprehensive understanding of the linguistic and phonetic bases of ASR system behaviour, which in turn will help courts and practitioners make best use of ASR systems in forensic casework.

1.2 Some challenges

ASR systems involve a complex set of processes to analyse and compare two speech signals (e.g., one of a known suspect and one of an unknown offender) in order to compute a numerical value which represents the strength of evidence. Features are initially extracted from the ‘voice-active’ portion of each speech signal (i.e. all parts of the signal containing speech). In state-of-the-art systems the features are then passed through a DNN to generate x-vectors. The x-vectors from two signals are compared to produce a score. Scores are numerical representations of the similarity and typicality of the x-vectors derived using pre-trained models within the system. However, scores are not directly interpretable, i.e. it is not always possible to judge whether a score is ‘high’ or ‘low’ in isolation, or given a single score, whether the system has made an error. In part this is due to differences between the conditions of the evidential comparison and the data used to train the system. Therefore, the final stage in the ASR processing chain is *calibration*, which is a means of interpreting the evidential score in light of same- and different-speaker scores from comparisons that are representative of the evidential comparison, but where the ground truth is known. This process of calibration converts the score to a likelihood ratio (LR); it is only at this point that the strength of evidence can be evaluated directly.

Speech recordings from forensic casework are also complex. Within a case, there is likely to be considerable variability across samples in terms of speaking style, interlocutor, and speaking level, as well as technical factors caused by e.g. channel differences, sample duration and background noise. Each forensic case is also unique, meaning that analysis could in principle involve speakers of any language, regional accent and socio-economic background, in any situational context.

1.3 A solution

In this study, we take an initial step towards understanding the sensitivity of ASR systems to different types of voices and the extent to which systems capture different phonetic information. We do this using highly controlled, and therefore forensically unrealistic, recordings of phoneticians in different vocal conditions. We focus on specific parts of the ASR chain in order to localise where speaker effects emerge and how they contribute towards system output.

Specifically, we focus on (i) x-vector speaker representations (extracted from a DNN using MFCCs) for different vocal conditions within-speakers, and (ii) the subsequent within-speaker score distributions which are generated by running a comparison in the ASR, but before any system calibration. All other variables are held constant (see description of data below) so that variation in the distributions can be interpreted relative to changes within a speaker's voice. While the focus on x-vectors and scores means that we cannot directly interpret whether differences are 'important' in terms of their ultimate impact on system performance, it does enable us to identify conditions of interest for further work.

2. DATA

This study uses a heavily controlled corpus which includes variation in speaker, vocal condition, time, and technical condition. In this paper, we report on a subset of the available material.

2.1 Participants

We report on data from two phoneticians: Paul Foulkes (PF - P1) and Francis Nolan (FN - P4). By using data from experienced phoneticians we could ensure there was minimal variability between participants when varying vocal conditions.

2.2 Vocal conditions

Each participant read the first two paragraphs of *The Rainbow Passage* in twenty four vocal conditions. These were selected to reflect changes in segmental and suprasegmental vocal parameters and fell into four groups (individual conditions are listed in Figures 1 and 2 below):

- *Modal voice.* This was used as a baseline for a speaker's 'normal' way of speaking.
- *Accent guises (5).* A range of linguistic changes, no significant shift in any one feature.
- *Miscellaneous (7).* Other forensically relevant variation.
- *Voice quality (7 supralaryngeal, 3 laryngeal).* One vocal setting at extreme, all others held as constant as possible.

Each vocal condition was repeated three times within a session (non-consecutive repetitions). Both participants took part in three sessions which were at least a week apart. In each session the conditions were repeated in the same order.

2.3 Technical conditions

Sessions were recorded in an anechoic chamber. Participants were seated at one end of the chamber throughout. Repetitions were simultaneously recorded in four technical conditions: headband microphone (DPA 4066 omnidirectional headset), near microphone (1m from participant), far microphone (2m from participant), and landline-to-VOIP call. Recordings were made in PCM WAV format with a 48kHz sample rate at 24 bits. For the purposes of the present study, only the headband microphone recordings were analysed

Individual repetitions of each vocal condition were extracted from within each session and are referred to throughout as a sample (i.e. 1 sample = 1 repetition of 1 condition in 1 session). There were 218 samples for PF, and 216 for FN. (There was one additional repetition of the high pitch and whisper conditions for PF.)

3. METHODS

3.1 The system

Analysis was carried out in the ASR software VOCALISE 2021 (version 3.0.0.1746) [2]. We report on outputs from the system at different stages in the comparison process.

3.2 x-vectors

For each participant, we first generated x-vectors for each sample using the default VOCALISE x-vector model.

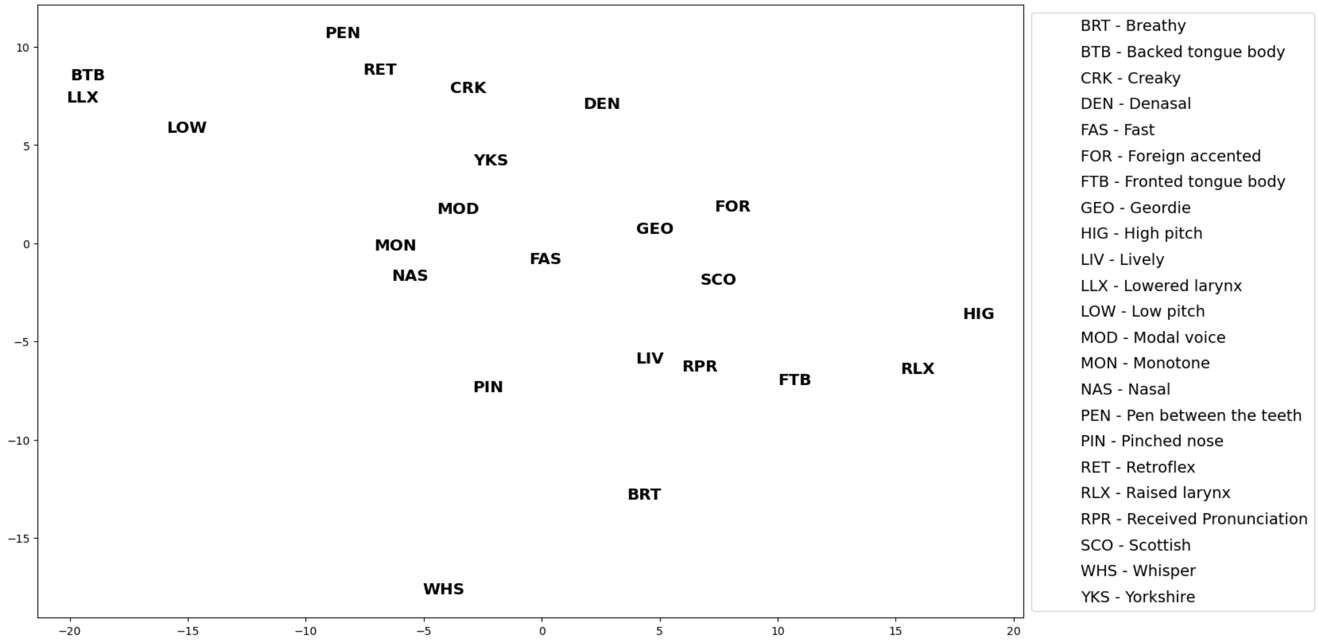

Figure 1: A t-SNE visualisation showing the average x-vector position for FN in each condition in 2-D space.

These were generated from MFCCs, which were extracted on a frame-by-frame basis across the sample and then passed through a DNN [2,8]. The x-vector is a representation of the speaker's voice in a given sample and contains 512 values. t-SNE (t-Distributed Stochastic Neighbour Embedding) plots were generated using the x-vectors to assess within-speaker patterns. t-SNE plots provide a way of visualising high-dimensional data in two or three dimensional space [9]; similar points appear closer together, and dissimilar points further apart [3].

3.3 Score distributions

Within-speaker comparisons were also carried out in VOCALISE using the default x-vector PLDA model. For each participant, each sample was compared to every other sample they produced. Each comparison results in a score. As noted earlier, scores are numerical representations of the similarity and typicality of the x-vectors derived using pre-trained models within the system, and interpretable only relative to other sets of similar scores. In this study we discuss only the within-speaker modal-to-other scores.

4. RESULTS

4.1 Within-speaker x-vector space

The t-SNE plot in Figure 1 visualises the x-vector space for FN (more detailed plots for both speakers are available online).

For FN, all within-condition repetitions clustered around the same area, so for ease of visualisation we have included only the average of the coordinates. The distribution of conditions in the x-vector space generally reflects the phonetic distance between the conditions. Low pitch, lowered larynx and backed tongue body all cluster in a similar area in the top left of the plot. At the bottom, breathy and whisper are separated from other conditions. Although his plot is not included here, PF's foreign accented repetitions did not cluster together, but instead three samples appeared away from the main cluster. These three were characterised by more marked larynx lowering, harsh phonation and some lingual retraction; this was in contrast to the other repetitions which were more breathy and perceptually higher in pitch. Taken together, it is clear that the x-vector distributions are capturing phonetic variability.

4.2 Within-speaker score distributions

The plot in Figure 2 shows the within-speaker (i.e. same speaker) modal-to-other score distributions. The dark grey distributions are PF, and the light grey are FN. The dashed line represents the median modal-modal score across PF and FN and is used here as a baseline for assessing which conditions demonstrate the biggest effects on scores. Individual medians for each distribution are shown as a solid line within the distribution. Vocal conditions are grouped, and ranked within each group based on the distance from modal (i.e. those which are more different appear at the bottom of the display).

The within-speaker modal-modal distributions for PF and FN overlap almost completely; both have similar degrees of variability across their repetitions in this condition and the scores are very similar.

The modal-accent guise scores are generally the closest to the modal-modal distributions. Both PF and FN varied their speech within these conditions on a range of different phonetic dimensions but the scores are generally in the same range as the modal-modal distributions. The exception to this is PF's foreign accented voice; as discussed earlier, these samples showed within-condition variability, and were also more different from his modal voice in a range of features.

Overall, the modal-supralaryngeal voice quality scores are the most different from the modal-modal scores. However, there is also considerable variability within this group, and between FN and PF. Modal-lowered larynx and modal-backed tongue body comparisons have distributions furthest from modal-modal, and FN and PF have different distributions, both in shape and degree of difference from modal.

For the miscellaneous conditions, those which involve supralaryngeal changes have the biggest impact on the score distributions. Modal-pinched nose, and modal-high pitch are most different from modal-modal. The difference between FN and PF in the modal-high pitch condition likely reflects the different strategies employed to produce high pitch; PF's high pitch involved some laryngeal raising and falsetto, thus involving supralaryngeal as well as laryngeal changes compared with PF's modal. The score distributions appear to reflect this difference.

When considering laryngeal voice quality, modal-whisper scores are the most different from those for modal-modal.

5. DISCUSSION

5.1 Within-speaker variability

The relationship between vocal conditions in the x-vector and score distributions reflects the degree of phonetic variability from modal. Supralaryngeal changes and whisper have the most divergent score distributions compared with the modal condition; the outcomes of these impact on a wide range of phonetic patterns (e.g. backed tongue body results in more retracted realisations of all vowels). In contrast, making intermittent, more targeted shifts to a range of features (e.g. the modal-accent guises) does not markedly impact the score distributions when compared to the modal-modal distributions.

5.2 Forensic applications

The findings allow us to further our understanding of potential mismatch conditions which might be relevant in casework situations.

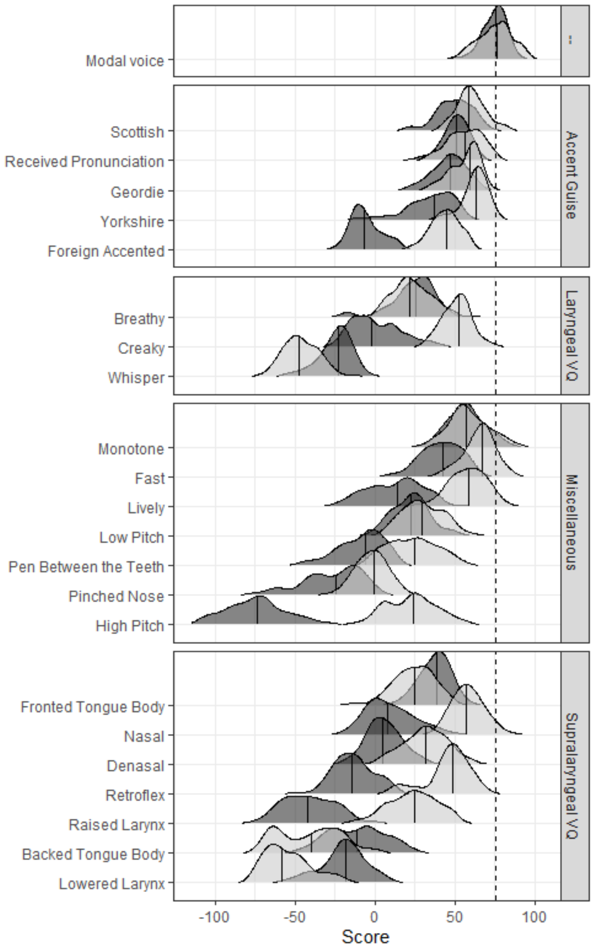

Figure 2: Within-speaker modal-to-other score distributions. Dark grey = PF; light grey = FN.

Although we do not know at this stage whether such factors would affect system performance. However, the fact that the system is sensitive to phonetic variation in a linguistically-predictable way (i.e. large supralaryngeal changes result in the most difference) at the two stages considered here is useful when it comes to addressing issues of explainability, for example, in a courtroom.

5.3 Future plans

In future work, we will consider further stages in the chain such as calibration and validation, and include different-speaker comparisons. We will also explore how the findings from this small controlled corpus can be extrapolated to other, less controlled data and a larger number of speakers.

6. ACKNOWLEDGEMENTS

This work was funded by ESRC grant ES/W001241/1.

7. REFERENCES

[1] Gold, E., French, P. 2019. International practices in

forensic speaker comparisons: second survey. *International Journal of Speech, Language and the Law* 26, 1–20.

[2] Kelly F, Forth O, Kent S, Gerlach L, Alexander A. 2019. Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. *AES International Conference on Audio Forensics.* Available: http://www.aes.org/e-lib/download.cfm?ID=20477

[3] Kelly, F., and Hansen, J. 2021. Analysis and Calibration of Lombard Effect and Whisper for Speaker Recognition. *IEEE/ACM Trans Audio Speech Lang Process* 29, 927–942.

[4] Morrison, G., Enzinger, E. 2016. Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction. *Speech Communication* 85, 119–126.

[5] Morrison, G., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., et al. 2016. INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International* 263, 92–100.

[6] Morrison, G., Enzinger, E. 2019. Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01)–Conclusion. *Speech Communication*. 112, 37-39

[7] R v Slade & Ors. 2015. Lord Justice Davis, Mr Justice Wilkie, Mr Justice Holroyde. Available: https://www.casemine.com/judgement/uk/5a8ff7a560d03e7f57eb0c31

[8] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[9] van der Maaten, L., Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 2579–2605.