Effects of vocal variation on the output of an automatic speaker recognition system

Vincent Hughes¹, Jessica Wormald¹, Paul Foulkes¹, Philip Harrison¹, Poppy Welch¹, Chenzi Xu¹, Finnian Kelly² and David van der Vloed³

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

Automatic speaker recognition (ASR) is increasingly used in forensic voice comparison cases. State-of-the-art systems utilise deep learning to convert acoustic features into compact speaker embeddings (e.g. x-vectors; Snyder et al. 2018). Embeddings from known and unknown voice samples are compared to generate a score, which in turn is calibrated to compute a numerical likelihood ratio (LR). Impressive performance of state-of-the-art systems has been reported with forensically realistic data (Morrison and Enzinger 2019), with marked improvements over previous generations of systems (e.g. i-vectors and GMM-UBM). Despite this progress, still relatively little is known about why certain voices perform better or worse within an ASR system, in part due to the abstract relationship between input and output, especially in state-of-the-art DNN-based systems. This issue is particularly important in the context of forensic voice comparison, where it is important for the practitioner to understand whether the output of an ASR system is reasonable given the input, and to explain the output to an end-user (e.g. a court).

In this study, we examine the effects of vocal variation on ASR output. We collected controlled recordings of six phoneticians reading the same text whilst systematically varying aspects of their speech production. Variations included modal voice, a range of laryngeal voice qualities and supralaryngeal vocal settings, high and low pitch, accent guises, and miscellaneous disguise techniques. Each speaker produced three repetitions of each vocal condition in each of three recording sessions, separated by at least one week. Analysis was conducted using the VOCALISE 2021 ASR system (version 3.0.0.1746; Kelly et al. 2019). X-vectors were generated for each sample from each speaker. Cross-session same-speaker (SS) and different-speaker (DS) comparisons were then conducted using PLDA to generate scores. Scores were converted to log LRs using calibration coefficients generated from condition-matched, cross-session SS and DS scores for 20 DyViS speakers (Nolan et al. 2009). Bayesian calibration with Jeffreys non-informative priors was used to account for the relatively small calibration set. Overall performance was evaluated using the log LR cost function ($C_{llr}$) and its two constituents: calibration loss ($C_{llr}^{cal}$) and discrimination loss ($C_{llr}^{min}$).

System performance was generally excellent across all matched-condition comparisons, with almost all vocal conditions producing $C_{llr}$s equivalent to the modal-modal condition. The exception was the whisper condition, which produced a markedly higher $C_{llr}^{cal}$ and a marginally higher $C_{llr}^{min}$. Unsurprisingly, condition mismatch had a much greater effect both in terms of calibration and discrimination loss. Whisper again had the largest effect on system output. In addition, vocal settings that substantially alter the supralaryngeal vocal tract (e.g. backed tongue body and lowered larynx) were found to have marked effects on system performance. Comparisons involving high pitch also generated relatively high $C_{llr}$ values (whereas low pitch did not), although interestingly this was most evident for speakers who achieved high pitch through modification of the vocal tract (e.g. through raising the larynx) rather than solely increasing the rate of vocal fold vibration. We discuss the implications of these findings for the use of ASR in forensic voice comparison casework.
References


