HASR → AHSR

(Human-assisted Automatic Speaker Recognition to Automatic-assisted Human Speaker Recognition)

Paul Foulkes, Peter French, Philip Harrison,
Vincent Hughes and Eugenia San Segundo
Department of Language and Linguistic Science, University of York, UK.
J P French Associates, York, UK.
paul.foulkes|peter.french|philip.harrison|vincent.hughes|@york.ac.uk
eugenia.sansegundo@csic.es

[Contribution for special session on ASR systems]

An international survey of forensic speaker comparison practises published in 2010 showed that 17% of laboratories/practitioners used automatic speaker recognition (ASR) technology in carrying out forensic voice comparisons (FVCs). More recent information from the manufacturer of a leading system indicates that the survey may substantially under-represent the true picture, in that, their system alone is currently in use for evidential purposes in c. 20 different countries. The rising popularity of ASR systems is unsurprising: speed of operation, testability to establish performance and error rates, relatively low dependency on the skills of the individual analyst and concomitant replicability of results make a compelling argument for their use in casework (French and Stevens, 2013).

However, ASR systems do make errors, even with high quality recordings. Although relatively small in number, these are almost exclusively false identifications rather than false rejections (French et al, 2009). From biological and linguistic perspectives, this bias is entirely predictable, and arises from limitations of the vocal tract as a biometric identifier. While the contours of the physical tract are probably unique to a speaker, its potential as a discriminant through MFCC analysis of its output is limited by (a) a lack of biological variation across individuals in terms of resonating chamber dimensions (Xue and Hao, 2006), and (b) the further reduction of such variation that does exist by linguistic socialisation - members of speech communities (language groups, accent groups) tend to converge in adopting similar supralaryngeal settings. In our view, these limitations are *inherent in the conceptual basis* of the ASR biometric and therefore not amenable to remedy by refinement of the algorithms or statistical models that ASR systems utilise. In light of this, it is proposed that we integrate ASR analysis into the battery of auditory- and acoustic-phonetic tests currently in place in many laboratories in order to ensure that errors produced by the systems are picked up by other forms of analysis (Gonzalez-Rodríguez et al, 2014). Thus, ASR testing is accorded a place alongside analysis of speech segments, rhythm, intonation, dysfluencies etc, in the FVC toolbox. In the Gold and French (2010) survey all those who used ASR claimed to provide some human input - ‘assistance’ - to the testing: human assisted automatic speaker recognition, or HASR. On the model proposed here the relative positions of the human and automatic elements are reversed: HASR → AHSR (automatic assisted human speaker recognition).

The main residual problem is how to incorporate the ASR results into an overall conclusion concerning (non-)identity of speakers. However, whilst there are correlations between MFCC measures as used by ASRs and long-term formant-based measures of supralaryngeal voice quality (LTFDs and LTFMs), there is no correlation with supralaryngeal profiles arrived at via an auditory-perceptual Vocal Profile Analysis (French et al, 2012) using the Laver scheme (Laver et al, 1981) The relative independence of ASR analysis from the main practical alternative for assessing supralaryngeal voice quality indicates that the results of the former may be fully taken into account in 'pitching' one's conclusion - however expressed - without running the risk of overestimating the strength of evidence by duplication of the same or similar information from different analytic methods and sources.

References

French, P., Foulkes, P., Harrison, P. & Stevens, L. (2012) Vocal tract output measures: relative efficacy, interrelationships and limitations. Paper presented at IAFPA 2012, Santander, Spain.

French, P., Harrison, P., Cawley, L., Bhagdin, A. & Clermont, F. (2009) Evaluation of the Batvox automatic speaker recognition system for use in UK based forensic speaker comparison casework. IAFPA conference, University of Cambridge.

French, P. & Stevens, L. (2013) Forensic speech science. Chapter Twelve of M. Jones & R. Knight (eds.) Bloomsbury Companion to Phonetics. London: Continuum.

Gold, E. & French, P. (2011b) International practices in forensic speaker comparison. International Journal of Speech, Language and the Law 18: 293-307.

Laver, J. et al. (1981) A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress* 14: 139-155.

Xue, S.A. & Hao, J.G. (2006). Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry. *Journal of Voice,* 20(3), 391-400.

Gonzalez-Rodríguez, J., Gil, J, Perez, R. & Franco-Pedroso, J. (2014) What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely-accepted trials. *Proceedings of Odyssey Speaker and Language Recognition Workshop*, Joensuu, Finland.