The complementarity of automatic, semi-automatic and auditory-phonetic measures of supralarygneal vocal tract output: an investigation based on speaker discrimination

Vincent Hughes[1], Philip Harrison[1,2], Paul Foulkes[1] and Peter French[2]
[1]Department of Language and Linguistic Science, University of York, UK
[2]J P French Associates, York, UK
{vincent.hughes|philip.harrison|paul.foulkes|peter.french}@york.ac.uk

A number of different methods can be used for characterising speakers in forensic voice comparison cases. These can broadly be grouped into three categories: (i) automatic, (ii) semi-automatic, and (iii) linguistic-phonetic methods. Automatic methods involve the extraction of, typically, cepstral features (e.g. Mel-frequency cepstral coefficients, MFCCs) from frames of equal length across the entire speech-active portion of a recording. Semi-automatic methods mimic the 'holistic', rather than segmental, approach of automatic systems, but typically extract more familiar, acoustic-phonetic features such as formant values from the vowel-only material (referred to as long term formant distributions; LTFDs). The linguistic-phonetic approach involves the componential analysis of features at potentially any linguistic level (segmental, suprasegmental, lexical, grammatical etc.) using a combination of auditory and acoustic methods. Within the field of forensic voice comparison there is now a growing move towards the integration of the best elements of the different approaches. However, a key issue is the extent to which they capture complementary speaker characterising information.

In this study, we examine long-term supralaryngeal vocal tract output measured in different ways: (i) automatic – MFCCs, (ii) semi-automatic – LTFDs, and (iii) linguistic-phonetic – supralaryngeal voice quality, analysed auditorily using Laver's Vocal Profile Analysis (VPA) protocol (Laver 1980). Data were extracted for 94 speakers from the Dynamic Variability in Speech (DyViS) corpus (Nolan et al. 2009) of young, standard southern British English males. MFCCs (incl. Δs and ΔΔs), LTFDs (incl. bandwidths and Δs), as well as Mel-weighted LTFDs ((M)LTFDs) were extracted from 20ms frames from the vowel-only portions of the recordings. Likelihood ratio-based speaker discrimination testing was conducted using the MFCCs in isolation, and in combination with LTFDS or (M)LTFDS. Generally, the addition of the LTFDs or (M)LTFDs did not improve the error rate (equal error rate; EER) compared with the MFCCs alone. This suggests that the formants are capturing essentially the same speaker characterising information as the MFCCs.

The best performing automatic system (EER = 3.23%) produced one false rejection (same speaker pair classified as different speakers) and 13 false acceptances (different speaker pairs classified as same speakers). These errors were examined in terms of the speakers' supralaryngeal VPA profiles. Of the 14 errors, nine involved speakers #67 and #72. Both of these speakers were found to have unremarkable VPA profiles. Further examination revealed a general correlation between the typicality of a speaker's VPA profile and the likelihood of producing an error, indicating that MFCCs do capture some of the same information as that encoded in auditory-based supralaryngeal VPA. Importantly, the errors were easily resolved using auditory analysis of laryngeal voice quality.

References

Laver, J. (1980) *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

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. *International Journal of Speech, Language and the Law* 16: 31-57.