The complementarity of automatic, semi-automatic, and phonetic measures of vocal tract output in forensic voice comparison

Vincent Hughes[1], *Philip Harrison*[1,2], *Paul Foulkes*[1], *Peter French*[1,2], *Colleen Kavanagh*[1], *and Eugenia San Segundo*[1]
[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|colleen.kavangh|eugneia.sansegundo}@york.ac.uk

In forensic voice comparison, automatic, semi-automatic and phonetic methods are available for evaluating voice evidence. Across the world, the phonetic approach is used predominantly in casework. This is due, in part, to the 'black box' perception of automatic systems and the lack of direct links between the features extracted and the underlying physiology. However, there is an increasing move towards the integration of the best elements of each approach (e.g. Gonzalez-Rodriguez et al., 2014). However, fundamental to the development of hybrid FVC systems is an understanding of the extent to which different methods capture complementary speaker-specific information.

In this study, we examine the potential improvement in the performance of a Mel-frequency cepstral coefficient-based (MFCC) automatic system with the inclusion of semi-automatic features (linear and Mel-weighted long term formant distributions; LTFDs and (M)LTFDs), and the role of auditory-based analysis of voice quality (VQ) in resolving *errors*. Recordings for 94 speakers from the DyViS corpus (Nolan et al., 2009) were analysed. Each sample was segmented into consonants and vowels using StkCV (Andre-Obrecht, 1988). The vowel-only portions of the samples were then divided into 20ms frames from which MFCC (12 MFCCs/12 Δs/12 ΔΔs), LTFD (F1~F4 frequencies/bandwidths/Δs), and (M)LTFD (Mel-weighted F1~F4 frequencies/bandwidths/Δs) feature vectors were extracted. VQ analysis was performed using a modified version of the vocal profile analysis (VPA) scheme (Laver, 1980; San Segundo et al., submitted). The 94 speakers were divided into development (31 speakers), test (31 speakers) and reference (32 speakers) sets. GMM-UBM likelihood ratios (LRs) were computed using the MFCCs, LTFDs and (M)LTFDs. The MFCC data were modelled with 1024 Gaussians, while 32 Gaussians were used for the formant data. Logistic-regression calibration and fusion was conducted using scores from the development data. Validity was evaluated using equal error rate (EER) and the log LR cost function (C_{llr}; Brümmer and du Preez, 2006).

The best performing MFCC system used MFCCs, Δs, and ΔΔs as input (EER=3.23%, C_{llr}=0.146). All of the LTFD and (M)LTFD systems performed considerably worse, with the (M)LTFD systems producing the poorest performance. For the LTFDs and (M)LTFDs, the addition of bandwidths and Δs did not improve performance. The fusion of LTFDs and (M)LTFDs with the MFCCs had essentially no effect on system performance, and in some cases validity got worse. Despite this, the best performing system overall used MFCCs+Δs+ΔΔs and LTFDs as input.

The *errors* – one false rejection and 13 false acceptances – produced by this system were evaluated in terms of VQ. A weak correlation was found between the typicality of a speaker's supralaryngeal VQ profile and the strength of evidence, with unremarkable speakers (i.e. those who were not distinctive in the group as a whole) more likely to produce weak or contrary-to-fact evidence. These results suggest that LTFDs, (M)LTFDs and supralaryngeal

VQ profiles capture some of the same speaker-specific information as MFCCs. However, the *error* pairs were still easy to separate based on auditory analysis, indicating that laryngeal VQ may provide independent complementary information which may improve the performance of (semi-)automatic systems.

References

Andre-Obrecht, R. (1988) A new statistical approach for automatic speech segmentation. *IEEE Transactions on Acoustics, Speech and Signal Processing* 36: 29-40.

Brummer, N. and du Preez, J. (2006) Application independent evaluation of speaker detection. *Computer Speech and Language* 20(2-3): 230-275.

Gonzalez-Rodriguez, J., Gil, J., Perez, R. and Franco-Pedroso, J. (2014) What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trials. In *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop*. Joensuu, Finland. pp. 33-40.

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

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.

San Segundo, E., Foulkes, P., French, J. P., Harrison, P., Hughes, V. and Kavanagh, C. (submitted) The use of the Vocal Profile Analysis for speaker characterisation: a methodological proposal.