Long term measures of the resonating vocal tract: establishing correlation and complementarity

Peter French, Paul Foulkes, Philip Harrison, Vincent Hughes, Eugenia San Segundo and Louisa Stevens

Department of Language and Linguistic Science, University of York, UK.

J P French Associates, York, UK.

{peter.french|paul.foulkes|philip.harrison|vincent.hughes|lcc108}@york.ac.uk

Underlying much of the research in forensic voice comparison (FVC) is the assumption that the vocal tract is a useful biometric for speaker discrimination and that individual differences in its anatomy and physiology will be reflected as speech resonances that are recoverable from its output. There are many ways in which the output of the tract may be observed and analysed, different methods deriving from both the linguistic-phonetic and automatic speaker recognition traditions. However, very little is known about the extent to which different vocal tract output measures are related, or about the potential for different measures to provide complementary information relevant to individual speaker characterisation. Building upon the findings of French et al. (2012), the research presented here arises from a project entitled Voice and Identity – Source, Filter, Biometric[1], and examines the correlations between features of multiple long-term automatic and semi-automatic measures of vocal tract output; namely mel-frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs), long-term formant distributions (LTFDs), long-term formant means (LTFMs), and auditory-based analysis of supralaryngeal vocal settings.

Data were extracted from the Task 2 studio recordings from the DyViS database (Nolan et al. 2009) for 100 male speakers of Standard Southern British English aged 18 – 25 years. First, supralaryngeal vocal settings were analysed auditorily by the sixth author (10% checked by the first author) using a modified version of the vocal profile analysis (VPA) scheme developed by Laver et al. (1981). To facilitate acoustic analysis, silences were removed from each recording using Praat. For consistency, the first four minutes of each sample was isolated. 16 MFCCs and LPCCs were then extracted from each 20 ms frame (hamming window) shifted at 10 ms intervals across the edited sample using the rastamat toolbox in MATLAB. Delta and delta-delta coefficients were appended to the features vector for each frame. For the LTF analyses, vowel-only samples were generated for each speaker from the edited recordings using the StkCV software. F1 to F4 measures were then extracted from 25 ms frames (Gaussians) shifted at 5 ms intervals using the iCAbS tracker (Harrison and Clermont 2012). The means of all values for each formant were then taken by-speaker to generate LTFMs. From these raw data, a Spearman correlation matrix was generated between each feature (e.g. individual cepstral coefficients) of each of the parameters analysed.

Strong correlations were revealed between LTFMs and individual VPA settings. Notably, F1 was found to be positively correlated with the raised larynx setting (r = 0.37; p = <0.01), while F2 was positively correlated with fronted tongue body (r = 0.27; p = <0.01). These are predictable given the articulatory bases underlying the acoustics of vowel formants. Considerably weaker correlations were found between the LTFDs (modelled using a GMM) and individual vocal settings. This is potentially due to the inclusion of information about

[1] Arts & Humanities Research Council, AH/M003396/1.

variability introduced by cross-phoneme formant analysis. No significant correlations were found between the more linguistically abstract cepstral coefficients and individual VPA settings. This lack of correlation potentially indicates that the speaker discrimination performance of ASR systems may be improved though the inclusion of complementary information derived from auditory-based VPA analysis. Further, this study also contributes towards improving our understanding of how the auditory categories of the VPA scheme relate to measurable acoustic output, and thereby provides a measure of acoustic validation of the scheme as a tool for use in FVC casework.

References

Andre-Obrecht, R. (1988) A new statistical approach for automatic speech segmentation IEEE Transactions on Acoustics, Speech and Signal Processing 36(1).

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.

Harrison, P. and Clermont, F. (2012) The influence of LPC order on the accuracy of formant measurements across speakers. Paper presented at IAFPA 2012, Santander, Spain.

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.

Nolan, F. et al. (2009) The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16(2): 31-57.