Voice quality analysis in forensic voice comparison: developing the vocal profile analysis scheme

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

Forensic voice comparison (FVC) is typically conducted using a combination of auditory- and acoustic-phonetic analyses. Among other features, voice quality (VQ) is considered especially valuable for distinguishing between speakers (Nolan 2005). In a recent survey of practitioners (Gold and French, 2011), 94% reported examining VQ, with 61% using a recognised framework such as Laver's VPA (Laver 1980). However, for several reasons these frameworks are not widely used for full systematic VQ analysis. Nolan (2005) lists some of the most common reasons for this: lack of training, practical considerations of time or quality of samples (e.g. telephone transmission or emotional speech). Besides, it is well known that analyses based on perceptual skills – even on those of trained phoneticians – are subject to bias and errors (Kent, 1997), which may call into question the reliability of such auditory evaluations. In relation to the evaluation of VQ in particular, inter-rater disagreements have plagued the history of perceptual studies (Kreiman and Gerratt, 2011). To the best of our knowledge, the inter-rater reliability of VPAs conducted using forensic material has not been reported, and the degree of within-speaker variability across multiple recordings has not been tested empirically. The present study addresses these issues and proposes developments for the VPA in FVC.

The modified VPA used by J P French Associates contains 32 supralaryngeal, phonatory and muscular tension settings. A setting is defined as the long-term tendency for some part of the vocal apparatus to adopt a particular configuration (Beck, 2005). The original protocol is very comprehensive and useful. However, it has also been criticised as too complex: "its greater scope is at the expense of reliability" (Webb et al. 2004, p. 429). The multidimensionality of VQ and the difficulties in isolating dimensions are widely reported (Kreiman and Gerratt, 2011).

Data consist of recordings (~ 5 minutes) of 100 young male RP speakers in the DyViS corpus (Nolan et al. 2009). High quality, near-end recordings of a telephone conversation from DyViS Task 2 were perceptually assessed by three analysts independently, using a modified version of Laver's VPA protocol. This had three rather than six scalar degrees for 'present' settings; degrees were collapsed and the highest degree, which would be considered pathological, was removed. Settings were scored on a 0 – 3 scale, where 0 means that a setting is absent (i.e. neutral) and values of 1 – 3 mean the setting is present to an increasing degree (noticeable/marked/extreme).

Inter-rater agreement across the three analysts was assessed using percentage agreement and several chance-corrected measures (Fleiss kappa). Preliminary results suggest a high degree of convergence across analysts, with relatively little disagreement about presence or absence of settings and typical differences of only one scalar degree for 'present' settings. The three analysts then held calibration and agreement sessions. These produced an agreed version of VPA ratings ready to be used in further research. This work involved investigating correlations between each of the 32 VPA settings. Empirical correlations were compared against predictions from phonetic theory to identify features which could be collapsed to further simplify the scheme. In future work we will compare VPAs established from Task 2 with those compiled from non-contemporaneous recordings (Task 1) of the

same speakers in order to evaluate within- and between-speaker variation. Pairs of same- and different-speaker VPAs are to be reduced to a Euclidean distance, and the distances used to generate likelihood ratio-like scores from which it is possible to generate system performance metrics. These results represent a step towards formalising VQ analysis in FVC, and simplifying the VPA scheme for wider forensic use. Further simplifications of this protocol are proposed based on the results of inter-rater agreement and correlation tests. Some simplification examples are: (1) reduction of supralaryngeal settings by merging those concerned with reduced or extended movements of different articulators; (2) reduction of correlated settings that result in sound effects that are difficult to distinguish perceptually (enlargement of the pharynx on lateral vs vertical dimensions – pharyngeal expansion ~ lowered larynx).

References

Beck, J. (2005). Perceptual analysis of voice quality: the place of Vocal Profile Analysis. In W.J. Hardcastle & J. Mackenzie Beck (Eds.) *A Figure of Speech: a Festschrift for John Laver. 285–322*. London/Mahwah, NJJ: Laurence Erlbaum Associates.

Gold, E. and French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech Language and the Law*, 18(2): 293–307.

Kent, R. D. (1997). Hearing and Believing: Some Limits to the Auditory-Perceptual Assessment of Speech and Voice Disorders. American Journal of Speech-Language Pathology, 5: 7–23.

Kreiman, J. & Gerrat, B. (2011). Comparing two methods for reducing variability in voice quality measurements, Journal of Speech, Language and Hearing Research, 54: 803-812.

Laver, J. (1980) The Phonetic Description of Voice Quality. Cambridge: CUP.

Nolan, F. (2005) Forensic speaker identification and the phonetic description of voice quality. In W.J. Hardcastle & J. Mackenzie Beck (eds.) *A Figure of Speech: A Festschrift for John Laver*. Mahwah NJ: Lawrence Erlbaum. pp. 385-411.

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(1), 31--57.

Webb, A.L.,· Carding, P.N., Deary, I.J., ·MacKenzie, K., Steen, N., Wilson, J.A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Otorhinolaryngology*, 261, 429–434.