Filled pauses as variables in speaker comparison: dynamic formant analysis and duration measurements improve performance for *um*

Sophie Wood, Vincent Hughes, and Paul Foulkes

Department of Language and Linguistic Science, University of York, York, UK.
{sophie.wood|vh503|paul.foulkes}@york.ac.uk

It is often hypothesised that filled pauses (FPs, i.e. *uh, um*) are useful variables in forensic speaker comparison (e.g. Künzel 1997, Tschäpe et al. 2005, Foulkes et al 2004, Jessen 2008). They offer several potential advantages over traditional segmental variables:

1. they are very frequent for most speakers and in most types of spontaneous speech;
2. they are typically longer than lexical vowels, and generally easier to measure;
3. they often abut silence, rendering them less susceptible to coarticulation, and thus in principle more consistent for the individual speaker;
4. there may be idiosyncratic patterns in the overall frequency of use, and in the discourse or syntactic contexts in which hesitations are used;
5. f0 patterns and durations may vary, as well as spectral components of vocalic elements;
6. the relative proportions of different FP types may also vary across speakers, i.e. whether speakers use vowel only (*uh*) or vowel+nasal (*um*) markers.

Here we present a study to investigate the discriminatory power of FPs, extending preliminary work presented by King et al (2013). FPs for 75 young male speakers of standard British English were analysed, drawn from Task 1 of the DyVis corpus (Nolan et al. 2009). The following acoustic properties were examined: 'static' midpoint frequencies of the first three formants in the vocalic portion; 'dynamic' measurements of the formants (i.e. quadratic curves fitted to 9 measurement points over the full vowel); and duration. Contemporaneous likelihood ratios were computed for independent sets of 25 development and 25 test speakers in MatLab (Morrison 2007) using Aitken & Lucy's (2004) Multivariate Kernel Density (MVKD) formula. Typicality was assessed using a reference set consisting of 25 speakers. Calibration coefficients were calculated based on the scores from the development data using a robust implementation of Brümmer's (2007) logistic regression procedure (Morrison 2009). The coefficients were then applied to the scores from the test data to generate calibrated log LRs. System performance was assessed using (i) Equal Error Rate (EER) as a metric of absolute discrimination between SS and DS pairs, and (ii) the log LR cost function (C_{llr}) (Brümmer & du Preez 2006), which provides a gradient assessment of system accuracy based on the magnitude of contrary-to-fact LRs.

Results are summarised in Table 1. For *uh* the static measurements outperform the dynamic measurements: EER is the same or slightly worse with the dynamic measurements, and C_{llr} is markedly worse in the dynamic measurement tests. This may be due to issues of overfitting trajectories that are essentially flat throughout the *uh* vocoid, meaning that static midpoints provide as much information without requiring so much input data. For *um*, on the other hand, dynamic measurements perform better than static measurements: EERs fall to less than 5% and C_{llr} reduces to less than 0.2. It is likely that the dynamic properties of *um* are more useful than those for *uh* because /VN/ FPs contain inherently more acoustic change between the vocalic and nasal portions. The addition of duration information further improves the EER and C_{llr} for *um.*

This study obtains LRs with EER scores below 5% using acoustic-phonetic features in spontaneous

speech recordings, which compares well with studies such as Becker, Jessen and Grigoras (2008). The study therefore strongly supports the view that FPs have excellent potential as variables in forensic speaker comparison cases, although formant dynamic data may only be useful for *um*, whereas static measurements provide equally good or better results for *uh.*

Table 1. Summary of results for *uh* and *um.*

Test:		**EER (%):**	**Cllr:**
Static	*Uh*	11.92	0.5246
	Um	11.92	0.3692
Static + duration	*Uh*	12.00	0.4876
(Static measurements fused with durations)	*Um*	8.92	0.2825
Dynamics	*Uh*	15.17	0.7068
	Um	4.67	0.1978
Dynamics + duration	*Uh*	11.92	0.7449
(Dynamic measurements fused with durations)	*Um*	4.17	0.1821

References

Aitken, C.G.G. & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics,* **54**, 109-122.

Becker, T., Jessen, M. & Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian Mixed Models. Paper presented at ISCA conference, Brisbane, Australia.

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

Foulkes, P., Carrol, G. & Hughes, S. (2004). *Sociolinguistics and acoustic variability in filled pauses.* Paper presented at IAFPA conference, Helsinki, Finland.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass,* **2**, 671-711.

King, J., Foulkes, P., French, P. & Hughes, V. (2013). Hesitation markers as a parameter for forensic speaker comparison. Paper presented at IAFPA conference, Tampa, Florida, USA.

Künzel, H.J. (1997). Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics,* **4**, 48-83.

Nolan, F., McDougall, K., de Jong, G. & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *The International Journal of Speech, Language and the Law*, **16**, 31-57.

Tschäpe, N., Trouvain, J., Bauer, D. & Jessen, M. (2005). *Idiosyncratic patterns of filled pauses.* Paper presented at IAFPA conference, Marrakesh, Morocco.