Hesitation markers as parameters for forensic speaker comparison

Jade King[1], *Paul Foulkes*[1,2,3], *and Peter French*[1,3]

[1]*Department of Language and Linguistic Science, University of York, York, UK.*
j_king18@hotmail.co.uk
[2]*New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, NZ.*
paul.foulkes@york.ac.uk
[3]*J P French Associates, York, UK.*
jpf@jpfrench.com

It is often hypothesised that vocalic hesitation markers are useful parameters for forensic speaker comparison (e.g. Shriberg 1994, Künzel 1997, Foulkes et al 2004, Tschäpe et al. 2005). Vocalic qualities of hesitation markers vary by language and dialect, but also idiosyncratically. As forensic parameters, hesitations have several potential advantages:

1. they are typically longer than lexical vowels, and thus the vocalic elements are generally easier to measure;
2. they may be preceded and/or succeeded by silence, thus rendering the vocalic elements less susceptible to coarticulation and thus in principle more consistent for the individual speaker;
3. there may be inter-speaker variation in the discourse or syntactic contexts in which hesitations are used;
4. f0 patterns and durations may vary, as well as spectral components of vocalic elements;
5. the relative proportions of different hesitation types may also vary across speakers, i.e. whether speakers use vowel only (*uh*) or vowel+nasal (*um*) hesitation markers.

Here we present an experimental study which aims to assess the relative power of the vocalic elements of hesitation markers for use in forensic case work.

Hesitation markers for 20 young male speakers of standard southern British English were analysed, drawn from Task 1 of the DyVis corpus (Nolan et al., 2009). Measurements were taken of the first three formants at the steady state midpoint of the vocalic portion, as well as overall duration. A total of 1965 tokens were analysed (1153 *uh*, 812 *um*; mean = 98 per speaker).

The formant and duration data were analysed separately for *uh* and *um*. Contemporaneous same-speaker (SS) and different-speaker (DS) Likelihood Ratios (LRs) were computed in MatLab (Morrison 2007) using Aitken and Lucy's (2004) Multivariate Kernel Density formula (MVKD). Separate analyses were conducted to include only formant data and then formant + duration data. Five speakers were used as test speakers, with remaining 15 as a background population.

To assess the relative performance of the hesitation data, LRs were also computed using formant data for three lexical vowels for the same speakers. (These data had previously been measured by Clermont REF.) A total of 1723 tokens were analysed for the three vowels KIT (542), DRESS (840) and TRAP (341). System performance was assessed using Equal Error Rate (EER) as a metric of absolute discrimination between SS and DS pairs, as well as the

log likelihood ratio cost function (C_{llr}) (Brümmer and du Preez 2006) which provides a gradient assessment of system accuracy based on the magnitude of contrary-to-fact LRs.

Results for all conditions are shown in Table 1. The hesitation data generally performed better than lexical vowels, as shown by the lower equal error rates (EERs) for both same speaker (SS) and different speaker (DS) comparisons. The C_{llr} values were also lowest for the hesitation data. Vowel+nasal hesitation markers performed better than vocalic *uh* in the SS condition. Adding duration data worsened performance.

Although the data set here is small, the analysis offers support for the hypothesis that hesitation markers are good variables to analyse in forensic speaker comparison. Further testing on a larger data set is merited. The addition of dynamic formant information might further improve the diagnostic power of the variable.

Condition		SS EER %	DS EER %	Cllr
Formants only	*Uh*	26	15	0.5669
	Um	0	23	0.5055
Formants and duration	*Uh*	60	45	0.8696
	Um	0	40	0.6200
Lexical vowels	KIT	38	28	0.8935
	DRESS	38	26	0.8303
	TRAP	0	33	0.5743

Table 1. Summary of results.

*May be of interest but having spoken to the Auckland people they were very impressed with non-calibrated Cllrs of less than 1 – assuming you can trust the results you've got here the Cllrs are very promising, but it could also be somewhat optimistic (or just wrong) because of the N speakers

References

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

Foulkes, P., G. Carroll & S. Hughes (2004). Sociolinguistics and acoustic variability in filled pauses. Paper presented at the IAFPA Annual Conference, Helsinki, Finland.

Künzel, H. (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. *Int. J. Speech, Lang. & Law* **16**, 31-57.

Shriberg, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California at Berkeley.

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