The individual and the system: assessing the stability of the output of a semi-automatic forensic voice comparison system

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

A body of research has shown the value of semi-automatic speaker recognition (SASR) based on long term formant distributions (LTFDs), with studies reporting equal error rates (EERs) of as low as 3% (Becker et al. 2008, Jessen et al. 2014). Much of this work has focused on matched conditions. Further, previous work has largely looked at system-level performance. However, a system with good overall validity may still have difficulties with the specific voices in a given forensic case. It is therefore important to understand more about the behaviour of individuals within systems and, ideally, identify 'problem' speakers prior to analysis (see Alexander et al. 2014, Ajili et al. 2016). In this paper we examine the effects of mismatched conditions on SASR performance and examine those speakers who are most and least sensitive to channel variation.

Task 1 and Task 2 recordings of 97 speakers from the DyViS corpus (Nolan et al. 2009) were analysed. The studio recording of Task 1 was used as the suspect sample. Four versions of the Task 2 recording were used as the offender sample: (1) near-end studio recording, (2) landline telephone recording, (3) high bit-rate (12.2kb/s) GSM recording and (4) low bit-rate (4.75kb/s) GSM recording. The GSM samples were created using the AMR Speech Codec Platform developed at the University of Auckland (see Alzqhoul et al. 2014). From each sample, 60 seconds of vowel material was used to extract F1, F2 and F3 frequencies, bandwidths and deltas (9 dimensional feature vector). Cross-validated, calibrated log likelihood ratios (LLRs) were computed using the GMM-UBM approach (Reynolds et al 2001) and logistic regression calibration. System validity was assessed using EER and C_{llr} (Brümmer and du Preez 2006).

Table 1. System performance for the four conditions (false hits/misses based on LLR threshold of 0)

	Suspect	*Offender*	*EER*	*False hits*	*Misses*	C_{llr}
	(Task1)	*(Task2)*	*(%)*	*(%)*	*(%)*	
(1)	HQ	HQ	10.33	11.58	8.24	0.37
(2)	HQ	Landline	25.95	31.06	17.53	0.73
(3)	HQ	GSM (HQ)	31.71	34.94	24.74	0.81
(4)	HQ	GSM (LQ)	31.99	34.19	28.87	0.83

Table 1 shows system validity across the four conditions. As predicted, there was a marked drop-off in performance for the mismatched conditions compared with the matched condition. The worst performance was found using the GSM samples, although there was no great effect of bit-rate. Within the system, individual speakers were found to behave differently. Figure 1 displays a zooplot (following Alexander et al. 2014) of the speakers with the highest and lowest means and standard deviations of LLRs produced across the four conditions. Voice quality was not found to be a predictor of an individual's position within the zoo-space. However, speakers who produced stronger evidence overall were found to be more variable in terms of their LRs. Speakers with high mean F3 also produced more LR variability. In this paper, we will expand on the reasons behind the behaviour of individual speakers within the system.

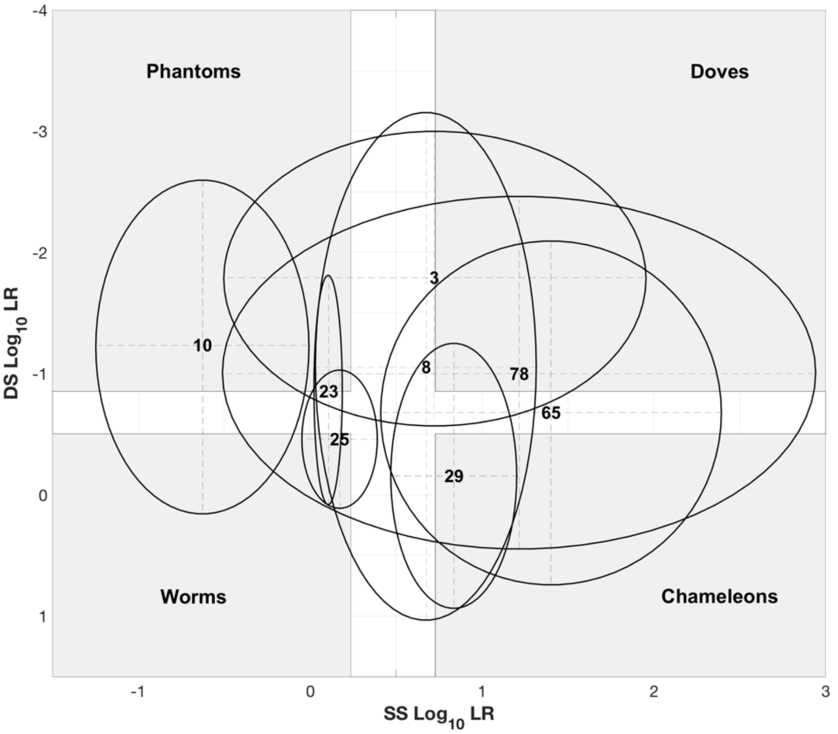

Figure 1. Zooplot of speakers with the highest and lowest means and standard deviations across the four conditions (speaker numbers are at the mean for each axis; ellipses are ±1 standard deviation)

References

Ajili, M. et al. (2016) Inter-speaker variability in forensic voice comparison: a preliminary evaluation. *Proceedings of ICASSP 2016*, 2114-2118.

Alexander, A., Forth, O., Nash, J. & Yager, N. (2014) Zooplots for speaker recognition with tall and fat animals. *Paper presented at IAFPA 2014*, Zurich, Switzerland. 31 August-3 September 2014.

Alzqhoul, E. A. S., Nair, B. B. T. & Guillemin, B. J. (2014) An alternative approach for investigating the impact of mobile phone technology on speech. *Proceedings of the Word Congress on Engineering and Computer Science*, Vol.1.

Becker, T., Jessen, M. & Grigoras, C. (2008) Forensic speaker verification using formant features and Gaussian mixture models. *Proceedings of Interspeech 2008*, 1505-1508.

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

Jessen, M., Alexander A. & Forth, O. (2014) Forensic voice comparisons in German with phonetic and automatic features using Vocalise software. *Proceedings of AES*, 28-35.

Nolan, F., McDougall, K., de Jong, G. & 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.

Reynolds, D. A., Qualtieri, F. & Dunn, R. B. (2001) Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10: 19-41.