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*[1,2]
Colleen Kavanagh[1], *Eugenia San Segundo*[1]

[1]Department of Language and Linguistic Science, University of York, UK
[2]J P French Associates, York, UK
{vincent.hughes|philip.harrison|paul.foulkes|peter.french}@york.ac.uk

Abstract

Semi-automatic systems based on traditional linguistic-phonetic features are increasingly being used for forensic voice comparison (FVC) casework. In this paper, we examine the stability of the output of a semi-automatic system, based on the long-term formant distributions (LTFDs) of F1, F2, and F3, as the channel quality of the input recordings decreases. Cross-validated, calibrated GMM-UBM log likelihood-ratios (LLRs) were computed for 97 Standard Southern British English speakers under four conditions. In each condition the same speech material was used, but the technical properties of the recordings changed (high quality studio recording, landline telephone recording, high bit-rate GSM mobile telephone recording and low bit-rate GSM mobile telephone recording). Equal error rate (EER) and the log LR cost function (C_{llr}) were compared across conditions. System validity was found to decrease with poorer technical quality, with the largest differences in EER (21.66%) and C_{llr} (0.46) found between the studio and the low bit-rate GSM conditions. However, importantly, performance for individual speakers was affected differently by channel quality. Speakers that produced stronger evidence overall were found to be more variable. Mean F3 was also found to be a predictor of LLR variability, however no effects were found based on speakers' voice quality profiles.

Index Terms: forensic voice comparison, semi-automatic speaker recognition, long term formant distributions, validity, biometric menagerie

1. Introduction

When presenting evidence to the courts, it is vital that forensic experts are able to explain their methods and procedures in an accessible way, such that they can be understood by lay people (i.e. jurors, judges and lawyers). This is essential if courts are to make reliable, informed, evidence-based decisions about the innocence or guilt of the accused. This issue is pertinent for forensic voice comparison (FVC) evidence, particularly when based on automatic speaker recognition (ASR) systems that can be perceived by the courts to be *black boxes* (see [1]). In many jurisdictions, experts continue to use linguistic-phonetic methods of analysis in FVC cases. A key benefit of analysing linguistic-phonetic features in FVC, and especially vowel formant frequencies, is that there is a mapping, albeit sometimes a non-linear one, between articulation and acoustic output [2,3]. They are also based on decades of well-understood and uncontroversial linguistic theory.

With these issues in mind, there has been an increasing focus on the use of semi-automatic systems (SASR) in FVC, reflected in the inclusion of SASR in the ENFSI methodological guidelines for best practice in 2015 [4]. SASR integrates linguistic and automatic methods: it combines (semi-)manual feature extraction (in contrast to automatic feature extraction, as in ASR systems), typically of formant frequencies (long term formant distributions; LTFDs), and automatic modelling, scoring and evaluation (as in normal ASRs). A growing body of research has shown that SASR based on LTFDs can be of considerable value. [5] reported an equal error rate (EER) of 4.14% based on F1, F2, F3, and F4 frequency values extracted from contemporaneous high quality studio recordings of British English speakers. [6] found comparable results for high quality studio recordings of German, with EERs ranging from 3% to 10.5%. Performance with more forensically realistic mobile-phone transmitted recordings has been shown to be slightly poorer, with [7] reporting an EER of between 4% and 18% depending on the input features and the number of Gaussians used to model those features.

This previous work has largely focused on matched (and typically good quality) technical conditions across suspect and offender samples. Yet mismatched conditions are the norm in forensic casework. For the majority of cases in the UK, the suspect sample is a relatively good quality recording of a police interview, while there can be considerable variability in offender samples (e.g. landline or mobile telephone recordings of varying qualities). Further, as with many studies in FVC and speaker recognition more generally, the focus of previous work on SASR has typically been on overall system performance (i.e. error rates). However, voices are complex and multidimensional. Speakers differ from each other as a result of a very wide range of features. Even if a system has very good validity overall, it may not perform particularly well for the specific evidential comparison in an individual case. In order to bridge the gap between research and casework, it is essential that we understand more about how individual speakers behave within our systems, rather than focusing exclusively on overall measures of validity. A small number of studies in FVC are beginning to address these issues (see [3,8]).

In the present study we assess the performance of a LTFD-based SASR system under matched and mismatched conditions as the quality of the offender recordings decreases – beginning with high quality recordings, and ending with more forensically relevant, low bit-rate GSM mobile recordings. A single set of speakers is tested in each condition using the same speech material. System performance (based on EER and C_{llr}) and the overall strength of evidence are analysed for each condition. The variability in output for individual test speakers is also evaluated over all conditions. Using the biometric menagerie and zooplots (see [9,10,11]) we identify which speakers are

more or less affected by the decrease in the channel quality of the offender sample, and attempt to identify the factors that may predict variability in performance.

2. Method

2.1. Materials

Recordings were drawn from the DyViS corpus of young male standard southern British English speakers [12]. Of the 100 available speakers, 97 were used, based on prior testing outlined in [13]. SASR testing was carried out under four different conditions according to the technical quality of the offender sample. Across all conditions, DyViS Task 1 was used as the suspect recording. Task 1 is a high quality, direct microphone recording (44.1kHz sampling rate, 16-bit depth), consisting of a mock police interview in which the participant is forced to lie about his involvement in a crime. DyViS Task 2 was used as the offender recording. Task 2 involves a telephone conversation between the participant and an accomplice to discuss a crime. The same speech material from Task 2 was used across conditions. Thus, there were no intrinsic, speaker-based differences (e.g. Lombard speech) between the recordings. Rather, the recordings represent a decrease in technical quality, and are reflective of the variety of recordings analysed in real forensic casework. The four versions of the offender samples used in this study are described below.

2.1.1. High quality (HQ)

The DyViS corpus contains the near-end, direct microphone recordings of the Task 2 telephone conversations. The original 44.1kHz recordings were downsampled to 10kHz.

2.1.2. Landline telephone (TEL)

The Task 2 speech was also recorded simultaneously at the far end of the telephone line, after transmission through a landline telephone network. The recordings had a sampling rate of 44.1kHz but were downsampled to 10kHz for the purposes of this study. The landline telephone transmission automatically imposes bandpass filtering on the signal of approximately 300Hz-3400Hz.

2.1.3. GSM Mobile with high (MOB_{HQ}) and low bit-rates (MOB_{LQ})

The high quality Task 2 recordings (§2.1.1) were manipulated to recreate 3G GSM mobile transmission – one of the most commonly used technologies for mobile transmission worldwide. Two sets of GSM recordings were created using different bit-rates to recreate high- and low-quality mobile transmission. The original recordings were initially resampled at a rate of 8kHz. The samples were then bandpass filtered, using a Hahn band filter, between 300Hz and 4000Hz to recreate typical mobile telephone filtering. The GSM AMR Speech Codec Platform [14] was used to apply the GSM codec to the recordings. This platform was used rather than passing the recordings through a real mobile network because the user has greater control over the settings. Using the AMR platform it is possible to change the bit-rate to a fixed value (whereas this is variable in the network depending on usage), as well as ensuring that frames are not dropped. The high quality GSM recordings were created using a bit-rate of 12.2kb/s (MOB_{HQ}) while the low quality recordings used a bit-rate of 4.75kb/s (MOB_{LQ}). Apart from bit-rate, the default settings were used (Version TS 26.073 with DTX disabled).

2.2. Preparation of recordings

The suspect recording (Task1) and the four versions of the offender recording (Task2) were prepared for analysis in the same way as described in §2.2 of [13]. This involved manual editing of recordings to remove non-speech sounds and overlapping speech, removal of sections containing clipping, voice activity detection to remove silences of greater than 100ms (using the *vadsohn* function in the VOICEBOX toolkit [15]), and segmentation of the signal into consonants and vowels using stkCV [16]. The first 60 seconds of vowel material were used for formant extraction. For each of the four versions of the offender sample, exactly the same 60 seconds of vowel material were used, in order to ensure that output was directly comparable.

2.3. Formant extraction

The 60 second audio samples were divided in 20ms frames with 10ms (50%) overlap between adjacent frames (6000 frames per sample). From each frame, F1, F2, and F3 frequencies and bandwidths were extracted using the Snack Sound Toolkit [17] with an LPC order of 12 and tracking four formants. Delta coefficients were also appended to the feature vector for each frame. Bandwidths and deltas were included in this SASR system as they have generally been shown to improve performance [6,13].

2.4. Conditions, system testing and evaluation

In this study, the following four conditions were tested:

(1) **Sus:** HQ (Task 1) vs. **Off:** HQ (Task 2)

(2) **Sus:** HQ (Task 1) vs. **Off:** TEL (Task 2)

(3) **Sus:** HQ (Task 1) vs. **Off:** MOB_{HQ} (Task 2)

(4) **Sus:** HQ (Task 1) vs. **Off:** MOB_{LQ} (Task 2)

For each condition, cross-validated same- (SS) and different-speaker (DS) scores were computed for all 97 speakers using the GMM-UBM [18] approach with MAP adaptation of means, variances and weights. The cross-validation involved retraining the UBM for each comparison, such that data from the comparison speaker(s) were not included in the UBM. In all cases, GMMs were fitted using 8 Gaussians based on performance in pre-testing. Score-level logistic regression calibration was then conducted, also using cross-validation [19]. Individual scores from each SS and DS comparison were calibrated individually, using a logistic regression model trained using all of the scores excluding those from comparisons involving the specific suspect and offender. This produced parallel sets of 97 SS and 4656 DS calibrated $\log_{10}$ likelihood ratios (LLRs) per condition. System validity was assessed using EER and the log LR cost function (C_{llr}) [20].

2.5. Analysing individuals

Individuals within the system were analysed in terms of the strength of the evidence they produced by calculating the means of LLRs for all of the SS and DS comparisons they were involved in across the four conditions. Variability in output was also assessed by calculating the standard deviations (SDs) of the SS and DS LLRs for each speaker across the four conditions. The behaviour of individual speakers is visualised in §3.2 using an adapted version of the zoo plot [10] described

in [11]. Zooplots are a way of visually representing different types of speakers within a system based on a typology called the biometric menagerie [9,10]:

- *Doves* are the best individuals for a biometric system, producing strong positive SS LLRs and strong DS LLRs
- *Sheep* are the majority of speakers who provide positive SS LLRs and negative DS LLRs, and thus well behaved
- *Worms* are the worst individuals, producing strong negative SS and strong positive DS LLRs
- *Phantoms* are successful at being separated from other speakers (strong negative DS LLRs) but struggle to be matched to themselves (strong negative SS LLRs)
- *Chameleons* match well to themselves but are poorly separated from others.

In §3.2, we fit multiple linear regression models to the SS and DS means and SDs. This is a means of testing which factors predict a speaker's position and variability in the zoo space. The SS and DS means and SDs were also used as independent variables (when not used as the dependent variable). Mean formant values for each speaker based on pooled data for their HQ Task1 and Task2 samples were also used as independent variables – we predicted that low mean F1 would result in higher LLR variability as it is more susceptible to the 'telephone effect' which artificially alters F1 values [21,22]. Auditory-based judgments of supralaryngeal and laryngeal voice quality (using data described in [13]) were also used as independent variables. The best model fit was identified using model comparison based on ANOVAs. A step-up approach was followed comparing the full model using all available predictors with combinations of fewer predictors.

3. Results

3.1. Overall performance

Table 1 shows system validity (EER and C_{llr}) for the four conditions tested.

Table 1: *Overall validity (EER and C_{llr}) for the four conditions using F1, F2, and F3 frequencies, bandwidths and deltas as input*

	Suspect	Offender	EER (%)	C_{llr}
(1)	HQ	HQ	10.33	0.37
(2)	HQ	TEL	25.95	0.73
(3)	HQ	MOB_{HQ}	31.71	0.81
(4)	HQ	MOB_{LQ}	31.99	0.83

As expected, condition (1) produced the best overall performance, achieving an EER of 10.33% and C_{llr} of 0.37. This is unsurprising given that this was the only matched-channel condition and used high quality studio recordings for both suspect and offender samples. This performance is slightly lower than the 6.45% (EER) and 0.255 (C_{llr}) reported for the same corpus using F1, F2, F3 and F4 in [13]. This suggests that F4 provides useful speaker discriminatory information and should be utilised in SASR systems where the channel characteristics of the samples allow.

A marked decrease in performance was found for the three mismatched conditions (2-4), relative to condition (1). Of these, condition (2) using the landline telephone offender sample produced the best validity (EER=25.95%, C_{llr}=0.73). The worst performance was found when using GSM mobile offender samples. Compared to condition (1), the performance of the GSM conditions was 20% worse in terms of EER and as much as 0.46 worse in terms of C_{llr}. Condition (4) using low bit-rate GSM offender samples produced the worst performance, although the difference between this and the high bit-rate GSM condition (3) was small. This suggests that SASR performance is relatively robust to GSM bit-rate.

3.2. Individuals

Figure 1 is a zooplot based on the SS and DS LLRs for all 97 speakers across the four conditions. The white cross in the middle of the zooplot is the interquartile range (the middle 50% of the data) of the SS and DS LLRs for all speakers. Within this region lie the *sheep* of the biometric menagerie, who produce solid positive SS LLRs and negative DS LLRs. The four shaded boxes represent the first and fourth quartiles of the LLRs. Speakers within these regions are classified as *phantoms*, *worms*, *chameleons* or *doves*. Although Figure 1 is based on all the LLRs, only those speakers with the highest and lowest mean SS and DS LLRs and the highest and lowest standard deviations (SD) are plotted. The speaker numbers (DyViS numbers) are at the mean SS and DS values, while the ellipses represent ±1 SD. Figure 1 highlights the heterogeneity in individual performance within the SASR system, both in terms of the strength of evidence and the variability of output across conditions. For instance, speakers #23 and 25 show little sensitivity to variation in channel across the four conditions, producing LLRs with low SS and DS SDs. Speaker #78, however, is extremely variable in terms of both SS and DS comparisons.

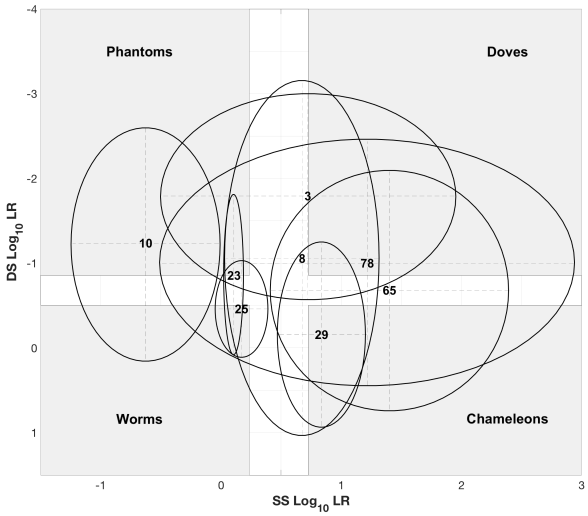

Figure 1: *Zooplot showing the speakers with the highest and lowest SS and DS means (DyViS speaker numbers) and standard deviations (ellipses = ±1SD)*

Linear regression models fitted to identify the factors that predict a speaker's position and variability within the zoo space revealed some significant effects. There was a significant correlation between SS and DS means (p<0.001). That is, speakers who produced strong positive SS LLRs generally also produced strong negative DS LLRs. Significant correlations were also found between means and SDs (p<0.001 for both SS and DS comparisons); the speakers producing the strongest LLRs were also the most variable. The only additional factor

predicting variability in LLRs was mean F3 (p<0.001). Speakers with higher mean F3 were found to have higher SDs for SS LLRs, but this was not significant for DS LLRs. No significant effects were found for F1 means or any voice quality features.

4. Discussion

Overall performance

The results in §3.1 show that the effects of channel mismatch on SASR performance can be considerable. The decrease in performance from the matched to the mismatched conditions is the difference between a system that provides relatively good speaker discrimination and systems that capture very little speaker-specific information. Landline telephone transmission produced performance closest to that of the matched condition (although still 16% worse in terms of EER and 0.36 worse in terms of C_{llr}). The poorest performance was found using the GSM samples. However, variation in bit-rate did not appear to affect performance substantially. This finding is consistent with [23] who report very little difference in C_{llr} values for low, medium and high quality GSM samples using MFCCs extracted from vowel phonemes as input. Interestingly, using the same segmental MFCC input, [24] found that GSM coding can lead to improvements in performance in matched conditions over un-coded speech. This suggests that, in our case, the cause of the drop in performance is due predominantly to mismatch, rather than quality in and of itself – although formant values may be affected differently to MFCCs.

Individuals

Across the four conditions tested, individual speakers were found to display different behaviour. Some produced very strong evidence, others weak evidence. Certain individuals were more variable, indicating a sensitivity to channel variation, while others produced much more stable LLRs. Speakers who produced strong SS LLRs were found to also produce strong DS LLRs. Otherwise, no significant effects were found to predict a speaker's position in the zoo space, and therefore classification in terms of the biometric menagerie. This suggests that neither mean F1, F2, and F3 values, nor any single auditorily-judged voice quality feature can be used to predict which speakers will perform well or badly within a formant-based SASR system. See [13] for a more systematic examination of the relationship between voice quality and LTFDs.

Some interesting effects were found in terms of the sensitivity of individuals to channel variation. Speakers who produced stronger LLRs were also found to be more variable. This finding is consistent with [25], who argued that variability in LLRs is greater where offender data lie at the tails of distributions (be that the suspect distribution or background distribution), since small changes to those distributions can have a dramatic effect on the probability of the evidence. As with strength of evidence, voice quality features were not found to predict a speaker's LLR variability. Although predicted, mean F1 did not correlate with LLR variability. Perhaps this is due to the fact that the samples from which the data were extracted contained a range of vowels, with values extending across the entire F1 range. This meant that speakers' mean F1 values were generally around 500Hz, and thus not especially susceptible to the 'telephone effect'. Although not tested here, the issue may be more due to the phonemic make-up of the sample analysed. We might predict that samples containing more close vowels (with inherently low F1) would produce more variable LLRs, than samples with more open vowels (with inherently higher F1).

However, speakers with high mean F3 values were more variable in terms of their SS LLRs. This may be due to the fact that we used the same default settings for formant extraction in the Snack Toolkit [17], with an LPC order of 12 and tracking four formants, for all speakers across all conditions. Although choosing system-level settings is the approach that has been followed in previous studies [5,6,7], it may have led to measurement issues here. Speakers with inherently high F3 are more likely to have F4 values close to or outside the upper bandpass threshold for telephone transmission. This is likely to cause F3 measurement errors, despite the fact that F3 itself isn't close to the bandpass threshold. This suggests, in line with [26], that it may be necessary to use channel- and speaker-specific (and possibly also vowel-specific) formant settings to help reduce the effect of channel mismatch and potentially improve overall system performance. This is something we intend to investigate in future work.

5. Conclusion

This paper has examined the effects of channel quality and mismatch on a formant-based SASR system. Mismatch was found to have an extremely detrimental effect on overall performance. We have also shown that there is considerable variability in individual behaviour both in terms of strength of evidence and sensitivity to the channel mismatch. The results highlight that analysis of individuals within forensic voice comparison (or indeed any biometric) systems should be an essential part of testing. It may also be that formant extraction settings need to be determined on a channel-by-channel and speaker-by-speaker basis. Only this information can allow the forensic scientist to understand whether their system is applicable to the voices under analysis in a given forensic case.

6. Acknowledgements

This research was funded via the UK AHRC grant Voice and Identity (AH/M003396/1).

7. References

[1] R v Slade & Ors [2015] EWCA Crim 71.

[2] P. Ladefoged and K. Johnson, *A Course in Phonetics (7th ed)*, Cengage Learning: Stamford (CT), 2014.

[3] J. Franco-Pedroso and J. Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition," *Speech Communication* vol. 76, pp. 61-81, 2016.

[4] A. Drygajlo et al., *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*, ENFSI, 2015.

[5] E. Gold, P. French and P. Harrison, "Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework," *Proc. Meetings on Acoustics* 19, 2013.

[6] T. Becker, M. Jessen and C. Grigoras, "Forensic speaker verification using formant features and Gaussian mixture models," *Proc. Interspeech*, pp. 1505-1508, 2008.

[7] M. Jessen, A. Alexander and O. Forth (2014) Forensic voice comparisons in German with phonetic and automatic features using Vocalise software," *Proc. AES*, pp. 28-35, 2014.

[8] M. Ajili, J. F. Bonastre, S. Rossatto and J. Kahn, "Inter-speaker variability in forensic voice comparison: a preliminary evaluation," *Proc. ICASSP*, pp. 210-217, 2016.

[9] G. Doddington, W. Liggett, A. Martin, M. Przybocki and D. A. Reynolds, "Sheeps, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," *Proc. Int. Conf. on Spoken Language Processing*, pp. 1351-1354, 1998.

[10] N. Yager and T. Dunstone, "The biometric menagerie," *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 32, pp. 220-230, 2010.

[11] A. Alexander, O. Forth, J. Nash and N. Yager, "Speaker recognition with tall and fat animals," *Paper at IAFPA*, 2014.

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

[13] V. Hughes, P. Harrison, P. Foulkes, P. French, C. Kavanagh and E. San Segundo, "Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-)automatic system testing," *Proc. Interspeech*, pp. 3892-3896, 2017.

[14] E. A. S. Alzqhoul, B. B. T. Nair and B. J. Guillemin, "An alternative approach for investigating the impact of mobile phone technology on speech", *Proc. World Congress on Engineering and Computer Science* vol. 1.

[15] VOICEBOX: Speech Processing Toolbox for MATLAB. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[16] R. Andre-Obrecht, "A new statistical approach for automatic speech segmentation", *IEEE Transactions on ASSP* vol. 36, pp. 29-40, 1988.

[17] K. Sjölander, Snack Sound Toolkit. http://www.speech.kth.se/snack/

[18] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing* vol. 10, pp. 19-41, 2000.

[19] N. Brümmer et al., "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST SRE 2006," *IEEE Transactions on Audio Speech and Language Processing* vol. 15, pp. 2072-2084, 2007.

[20] N. Brümmer, and J. du Preez, "Application-Independent Evaluation of Speaker Detection," *Computer Speech and Language* vol. 20, pp. 230-275, 2006.

[21] H. Künzel, "Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies," *IJSLL* vol. 8, pp. 80-99, 2001.

[22] C. Byrne and P. Foulkes, "The 'mobile phone effect' on vowel formants," *IJSLL* vol. 11, pp. 83-102, 2004.

[23] E. A. S. Alzqhoul, B. B. T. Nair and B. J. Guillemin, "Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison," *Science and Justice* vol. 55, pp. 363-374, 2015.

[24] B. B. T. Nair, E. A. S. Alzqhoul and B. J. Guillemin, "Impact of various GSM network factors on forensic voice comparison," *Proc. Australasian Int. Conf. Speech Science and Technology*, pp. 137-140, 2016.

[25] V. Hughes, *The Definition of the Relevant Population and the Collection of Data for Likelihood Ratio-Based Forensic Voice Comparison*, Unpublished PhD Thesis, University of York, UK, 2014.

[26] G. K. Vallabha and B. Tuller, "Systematic errors in the formant analysis of steady-state vowels," *Speech Communication* vol. 38, pp.141-160, 2002.