EFFECTS OF FORMANT ANALYSIS SETTINGS AND CHANNEL MISMATCH ON SEMI-AUTOMATIC FORENSIC VOICE COMPARISON

Vincent Hughes[1], Philip Harrison[1,2], Paul Foulkes[1], Peter French[1,2], Amelia J. Gully[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|amelia.gully}@york.ac.uk

ABSTRACT

This study examines the sensitivity of formant-based semi-automatic speaker recognition systems to feature extraction settings and channel mismatch. A total of 200 systems were tested, varying LPC order and the maximum number of formants tracked across four channels: studio quality, landline telephone, and two GSM mobile phone samples with different bit-rates. For each system, calibrated log likelihood ratios were computed for 97 speakers using formants extracted from 60 seconds of vowel-only material. System performance was affected markedly by formant settings, with EER ranging from 8% to 37% and C_{llr} ranging from 0.28 to 0.88 in the high quality condition. However, some individuals are more sensitive to such variation, meaning that system performance is entirely dependent on the specific speakers tested. This issue is discussed in light of the ongoing debate about the validation of methods and testing of systems under the conditions of the case.

Keywords: forensic voice comparison, formants, validity, LPC order

1. INTRODUCTION

Within the field of forensic voice comparison there is a growing trend towards integrating methods from speech technology (i.e. automatic speaker recognition; ASR) and linguistics and phonetics. Historically these fields developed separately, but there is now growing evidence that integrated approaches can improve overall system performance and help us better understand the behaviour of systems under different conditions [8,10,13]. This is especially important in the context of forensic evidence, which must be explained to and understood by non-specialist end-users (e.g. juries, police).

There has been increasing focus on the speaker discriminatory potential of semi-automatic speaker recognition (SASR) systems, which represent a hybrid between ASR and linguistic-phonetic methods [11]. SASR involves the extraction of linguistic-phonetic acoustic features, typically formant frequencies (long-term formant distributions; LTFDs). Formants are extracted only from the vowel material within a sample and so some degree of pre-processing is required. This contrasts with ASR systems that extract features from the entire speech-active portion of a sample. Modelling, scoring and evaluation in SASR are performed automatically.

With good quality, or at least matched-quality materials, SASR systems have been shown to perform well, with studies reporting equal error rates (EER) as low as 3-4% [3,9]. [14] provide the first systematic analysis of channel mismatch on LTFD-based SASR systems. They examined system performance and the sensitivity of log likelihood ratios (LLRs) for individual speakers to channel variation. Predictably, the best performance was found using high quality studio samples, producing an EER of 10% and a log LR cost (C_{llr}; [4]) of 0.37. Performance degraded considerably with mismatched recordings, with the comparison between high quality suspect samples and low bit-rate GSM mobile phone offender samples producing an EER of 32% and a C_{llr} of 0.83. Despite variation in system performance, LLRs for some speakers were found to be stable across conditions, while other speakers were much more sensitive to channel mismatch. The only predictor of sensitivity was mean F3, such that speakers with high mean F3 produced more variable LLRs. It was suggested in [14] that this may be due to measurement error issues.

Previously, SASR systems have been tested with a single configuration of formant extraction settings which are applied to all speakers across all conditions, with little or no post-processing applied to deal with measurement errors. However, it is well known that formant measurements are sensitive to the settings and software used [12] and to channel variation [6], and that caution should be exercised over using unchecked data [7]. The present study builds on [14] to examine the effects of formants settings (LPC order and the number of formants tracked) on both the validity of SASR systems and the LLRs produced for individual speakers. This study uses the same materials across the same conditions as [14]: high quality studio samples, landline telephone samples, and two GSM mobile phone samples with high and low bit-rates. Output is compared in terms of overall system performance, as well as validity metrics for individual speakers.

2. METHOD

2.1. Materials

Analysis was performed using recordings from 97 male, standard southern British English speakers aged 18 to 25 from the DyViS corpus [16]. Participants were recorded in a mock police interview (Task1) and a telephone conversation with an accomplice (Task2). High quality, studio recordings of Task1 were used as suspect samples. Task2 was used as the offender sample. Four versions of Task2 were tested:

(1) *High quality (HQ)*: To capture optimal performance in matched conditions, the near-end, studio recordings were used.
(2) *Landline telephone (TEL)*: These were recordings made at the far-end of a landline telephone line (downsampled to 8kHz).
(3) *GSM mobile with high bit-rate (MOB_{HQ})*: 3G GSM samples were created by down-sampling to 8kHz and bandpass filtering between 300Hz and 3400Hz. The GSM codec was then applied using the AMR Speech Codec Platform [2], which allows the user to specify bit-rate and frame loss settings. In the MOB_{HQ} condition, a fixed bit-rate of 12.2kb/s was used.
(4) *GSM mobile with low bit-rate (MOB_{LQ})*: The same procedures as above were applied, but with a fixed bit-rate of 4.75kb/s.

2.2. Preparation of recordings

Samples were automatically divided into consonants and vowels using stkCV [1]. For each speaker, 60 seconds of vowel material was used (see [13,14]).

2.3. Formant extraction

Formants were extracted from the vowel-only samples using a 20ms window with 10ms shift. From each frame, the first three formant frequencies and bandwidths were extracted using the Snack Sound Toolkit [19] with deltas (capturing dynamic frame-to-frame variation) also appended. Different sets of formant data were extracted using different settings: based on LPC order (ranging from 12 to 16) and the number of formants tracked (3 or 4).

2.3. Testing and evaluation

Four transmission conditions were tested in this study: (i) HQ-HQ, (ii) HQ-TEL, (iii) HQ-MOB_{HQ}, and (iv) HQ-MOB_{LQ}. Within each condition, three variables relating to formant settings were analysed: suspect LPC order, offender LPC order, and the number of formants tracked. This produced a total of 200 systems (50 systems in each condition) – the term *system* here is used to describe a configuration of settings, rather than the more specific use of the term to describe an ASR system.

For each system, cross-validated same- (SS; 97) and different-speaker (DS; 4,656) scores were computed using GMM-UBM [16] with MAP adaptation of means, variances, and weights. Cross-validation was performed by retraining the UBM for each comparison, excluding data from the suspect and offender being compared each time. GMMs were fitted using eight Gaussians (based on pre-testing). Scores were then converted to LLRs using cross-validated logistic regression [5]. This involved calibrating each score individually, training the logistic regression model on scores from comparisons that did not involve the suspect or offender. This produced parallel sets of calibrated LLRs that were used to calculate the EER and C_{llr} for each system. For both metrics, the closer the value is to zero the better the performance.

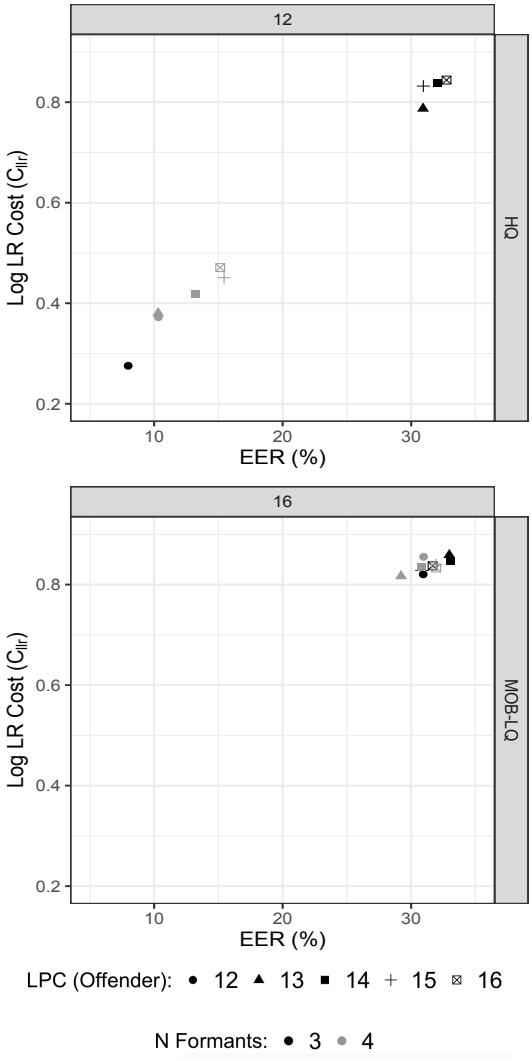
The effects of channel and formant settings were also assessed in terms of individual speakers. Within each transmission condition, the SS and DS LLRs from comparisons involving each speaker as suspect or offender were extracted and used to calculate a speaker-specific EER and C_{llr} – with 50 systems in each condition, each speaker produced 50 SS LLRs and 4,800 DS LLRs. Speaker-specific validity is used here as a measure of a speaker's sensitivity to changes in formant settings (i.e. the higher the EER or C_{llr} the more sensitive a speaker is to changing formant settings). Importantly, the speaker-specific validity metrics also capture the inherent discriminatory potential of a speaker, and this interacts with their sensitivity to formant settings. To account for this, speaker-specific validity is examined in light of the LLRs produced by individuals in the best performing condition (in this case, HQ-HQ, Suspect LPC: 12, Offender LPC: 12, tracking 3 formants).

3. RESULTS

Figure 1 displays two sets of system validity results. These were chosen because they were the conditions that produced the most and least variability in terms of EER and C_{llr}. Figures showing the performance of all 200 systems can be found here: [https://vincehughes.files.wordpress.com/2019/03/allsystems.pdf]. The greatest variability was found in the HQ-HQ condition using a suspect LPC order of 12 (which is, importantly, the default in Snack), with EERs ranging from 7.9% to 32.8% and C_{llr}s ranging from 0.28 to 0.84. This, more than any other condition, highlights the potential magnitude of the effects of formant settings on SASR performance.

While overall speaker discriminatory power was generally much lower in the mismatched conditions (HQ-TEL, HQ-MOB_{HQ}, and HQ-MOB_{LQ}), system validity was more robust against changes in formant settings. As shown in Figure 1(b), in the HQ-MOB_{LQ} condition, EERs ranged from 29.2% to 33.1% and C_{llr}s ranged from 0.82 to 0.86. This is likely due to the fact that there is much less potential for variability in system performance with degraded recordings that inherently produce poorer systems. Even with accurate formant measurements the HQ-MOB_{LQ} condition will never produce EERs as low as the 7.9% in the HQ-HQ condition.

Figure 1: System validity (EER and C_{llr}) in the most (top: HQ-HQ, Suspect LPC = 12) and least (bottom: HQ-MOB_{LQ}, Suspect LPC = 16) variable conditions



No systematic patterns were found in terms of the effects on performance of LPC order or the number of formants tracked. For some combinations of suspect and offender LPC orders, tracking three formants consistently outperformed tracking four formants (e.g. suspect LPC order of 13 in the HQ-HQ condition), while for other combinations four formants were better (e.g. suspect LPC order of 12 in the HQ-MOB_{LQ} condition). This suggests that changing formant settings does not have a uniform effect on the system as a whole. Rather, different settings are better for certain speakers.

Figure 2: Speaker-specific system validity (EER and C_{llr}) in the four transmission conditions (each dot represents a single speaker and is calculated using their SS and DS LLRs across all settings)

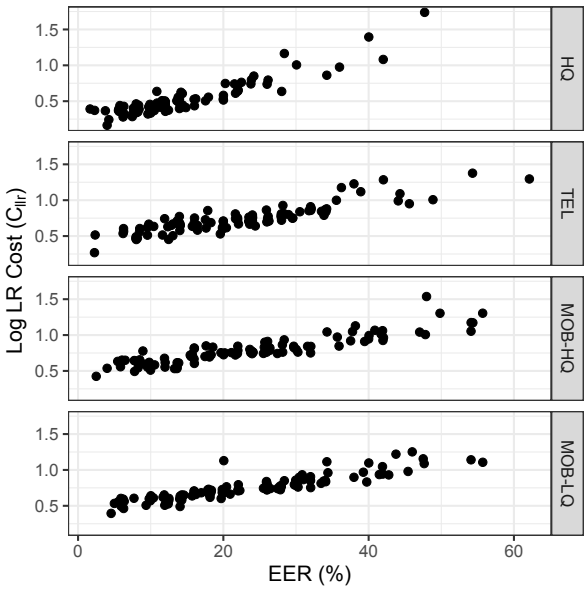


Figure 2 shows speaker-specific validity in the four transmission conditions. To assess these patterns, a C_{llr} of 1 was used as a threshold to determine performance. A system that consistently produces LLRs of 0 (i.e. provides no useful speaker discriminatory information) will have a C_{llr} of 1. Thus, a C_{llr} of less than 1 means that a system is capturing speaker discriminatory information, while a C_{llr} of more than 1 means very poor performance. It is clear from Figure 2 that the proportion of speakers producing C_{llr}s of more than 1 is greater in the mismatched conditions (up to 15.46% in the MOB_{HQ} condition) compared with the matched, HQ-HQ condition (5.15%). Given the results of overall system testing, this is likely due to reduced speaker-discriminatory power in the mismatched conditions, rather than greater sensitivity of speakers to changes in formant settings. However, a number of speakers recurrently produced high C_{llr}s across conditions. Notably, speaker #10 (DyViS code) produced C_{llr}s over 1 in all four conditions, while speakers #32, #44, #58, #103, and #111 produced C_{llr}s over 1 in three of the four conditions. Examining these speakers in more detail reveals some interesting patterns. For some speakers (#32), the high C_{llr}s are due to the fact that they are inherently difficult to match with themselves, and to discriminate from others. Other speakers (#58, #103, #111) produced strong SS and

DS LLRs in the optimal condition and so the high C_{llr}s directly reflect sensitivity to formant settings. These are, in a way, the most problematic speakers in terms of choosing formant settings, since changes to settings have a dramatic effect on their LLRs, leading them to produce more high magnitude errors. Finally, some speakers (#10, #44) display an interaction between discriminatory power and sensitivity to formant settings. These speakers generally produce average to weak LLRs, but are by no means the worst performing speakers in the system. Since their LLRs are inherently closer to the threshold, even small variation in formants as a result of the settings used can have a substantial effect on the proportion and magnitude of errors.

4. DISCUSSION

4.1. Overall performance

Our results show that SASR performance is heavily dependent on the formant settings used. This is the case even, and in fact most notably, in high quality conditions where the overall speaker discriminatory power is likely to be higher. The results serve as a reminder to exercise caution when interpreting the performance of an SASR system which has only been tested with a single set of settings. At the very least, SASR systems should be tested with different settings to assess the best configuration for the comparison in a case. This would also provide a means of understanding the potential range of results the SASR system could produce with different settings.

4.2. The role of the individual in system testing

The lack of systematic patterns show that individuals are not affected equally by changes in formant settings across transmission conditions. Some individuals are, on the whole, more or less sensitive to changes in formant settings. Other individuals are likely to be sensitive only to specific changes in formant settings.

The issue of the role of individual speakers within a system is important in the context of the wider debate about the *paradigm shift* across forensic science and calls for the testing of systems under the conditions of the case at trial. Morrison [15] states that "the test data must be sufficiently representative of the relevant population and sufficiently reflective of the speaking styles and recording conditions in the case." It is, of course, useful to know how well your system performs generally with such test data. However, ultimately, the issue is that the analyst needs to know the probability of having made an error (i.e. producing a contrary-to-fact LR) for the specific suspect and offender in the case, rather than the general performance of the system. Using test data of the type described by Morrison [15] does not necessarily answer this question.

We would argue that it is more critical to understand the behaviour of individual speakers, or *types* of speakers, within the system, and thus be able to predict how specific suspect and offender voices may perform. This requires knowledge of the factors that alter speaker behaviour in a system. Clearly, as this study has shown, formant settings are one such factor. However, the broader questions of why speakers are more or less sensitive to changes in formant settings and why this affects speaker discrimination are more difficult to answer. It is well known, to those who often measure formants manually, that some speakers' formants are much harder to track than others. The potential reasons for this are likely numerous and complicated. Phonation (laryngeal voice quality) is one potential explanation. Speakers with habitually non-modal phonation may produce more formant errors since they provide a poorer fit to the LPC model. However, we found no consistent pattern when analysing the voice quality profiles (from [18]) of the problematic speakers identified at the end of section 3. The variability in formant measurements may also be due to filter configurations that generate wide bandwidths and, thus, less prominent peaks, as well as natural within-speaker variability in the voice. However, more research is required to explore these issues.

5. CONCLUSIONS

This study highlights the importance of formant settings across different transmission conditions in testing SASR systems. The results show that individuals are sensitive to these factors to different extents, and that this sensitivity interacts with speaker discriminatory potential to affect the resulting LLRs. The results here are also relevant to the wider phonetic community. In line with [7], caution should be exercised using automatically generated formant data, especially in the context of large-scale, corpus-based studies that rely on the scale of data to reveal subtle effects; e.g. related to predictability. Given our findings, we consider it preferable for any phonetic study using formant measurements to minimally use speaker-specific settings. However, as suggested by [12] it may be that vowel-specific settings are also necessary.

6. ACKNOWLEDGEMENTS

This research was funded by the Arts & Humanities Research Council (AHRC) project *Voice and Identity* (AH/M003396/1).

7. REFERENCES

[1] Andre-Obrect, R. 1988. A new statistical approach for automatic speech segmentation. *IEEE Transactions on ASSP* 36, 29–40.

[2] 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. *Proc. World Congress on Engineering and Comp. Science* 1, San Francisco.

[3] Becker, T., Jessen, M., Grigoras, C. 2008. Forensic speaker verification using formant features and Gaussian mixture models. *Proc. Interspeech*, Brisbane, 1505–1508.

[4] Brümmer, N., du Preez, J. 2006. Application-independent evaluation of speaker detection. *Comp. Sp. Lang* 20, 230-275.

[5] Brümmer, N., Burget, L., Černocký, J., Glembek, O., Grétzl, F., Karafiát, M., van Leeuwen, D.A., Matějka, P., Schwarz, P., Strasheim, A. 2007. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST SRE2006. *IEEE Transactions on Audio Speech and Language Processing* 15, 230–275.

[6] de Decker, P., Nycz, J. 2011. For the record: which digital media can be used for sociophonetic analysis? *U. Penn Working Papers in Linguistics* 17(2), 51–59.

[7] Foulkes, P., Docherty, G., Shattuck-Hufnagel, S., Hughes, V. 2018. Three steps forward for predictability: consideration of methodological robustness, indexical and prosodic factors, and replication in the laboratory. *Linguistics Vanguard* 4(2). doi: 10.1515/lingvan-2017-0032.

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

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

[10] Gonzalez-Rodriguez, J., Gil, J., Pérez, R. Franco-Pedroso, J. 2014. What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trials. *Proc. Odyssey*, Joensuu, 33–40.

[11] Greenberg, C., Martin, A., Brandschain, L., Campbell, J., Cieri, C., Doddington, G., Godfrey, J. 2010. Human assisted speaker recognition in NIST SRE2010. *Proc. Odyssey,* Brno, 180–185.

[12] Harrison, P. 2013. *Making Accurate Formant Measurements: An Empirical Investigation of the Influence of the Measurement Tool, Analysis Settings and Speaker on Formant Measurements*. Unpublished PhD Thesis, University of York.

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

[14] Hughes, V., Harrison, P., Foulkes, P., French, J.P., Kavanagh, C., San Segundo, E. 2018. The individual and the system: assessing the stability of the output of a semi-automatic forensic voice comparison system. *Proc. Interspeech*, Hyderabad, 227–231.

[15] Morrison, G.S. 2018. Admissibility of forensic voice comparison testimony in England and Wales. *Criminal Law Review* 1, 20–33.

[16] Nolan, F., McDougall, K., de Jong, G., Hudson, T. 2009. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *IJSLL* 16, 31–57.

[17] Reynolds, D. A., Qualtieri, T. F., Dunn, R. B. (2001) Speaker verification using adapted Gaussian Mixture Models. *Digital Signal Processing* 10, 19–41.

[18] San Segundo, E., Foulkes, P., French, J.P., Harrison, P., Hughes, V., Kavanagh, C. 2018. The use of the Vocal Profile Analysis for speaker characterisation: methodological proposals. *JIPA*. doi: 10.1017/S0025100318000130.

[19] Sjolander, K. 2005. Snack Sound Toolkit (v.2.2.10). http://www.speech.kth.se/snack/