Formant dynamics and durations of *um* improve the performance of automatic speaker recognition systems

Vincent Hughes, Paul Foulkes, Sophie Wood

Department of Language and Linguistic Science, University of York, UK

vincent.hughes@york.ac.uk, paul.foulkes@york.ac.uk

Abstract

We assess the potential improvement in the performance of MFCC-based automatic speaker recognition (ASR) systems with the inclusion of linguistic-phonetic information. Likelihood ratios were computed using MFCCs and the formant trajectories and durations of the hesitation marker *um,* extracted from recordings of male standard southern British English speakers. Testing was run over 20 replications using randomised sets of speakers. System validity (EER and C_{llr}) was found to improve with the inclusion of *um* relative to the baseline ASR across all 20 replications. These results offer support for the growing integration of automatic and linguistic-phonetic methods in forensic voice comparison.

Index Terms: forensic voice comparison, automatic speaker recognition, hesitation markers, formant dynamics

1. Introduction

Forensic voice comparison (FVC) accounts for the majority of casework conducted by forensic speech scientists. FVC typically involves the comparative analysis of speech samples of a known suspect (e.g. police interview) and an unknown offender (e.g. covert drug deal). In such cases, it is the role of the expert to evaluate the strength of the speech evidence under the competing propositions of the prosecution (i.e. the suspect and the offender are the same person) and the defence (i.e. the suspect and the offender are different people).

Two sets of methods are commonly used in FVC: auditory-acoustic (linguistic-phonetic) analysis and automatic speaker recognition (ASR). These methods have largely developed independently. However, a growing body of research focuses on the integration of the methods to improve the performance of FVC systems. [1] and [2] investigated the performance of a generic Mel frequency cepstral coefficient (MFCC)-based ASR system when fused with formant and tone (f0) trajectories of vowels in standard Chinese. The results show that the fusion of linguistic-phonetic and ASR systems improves performance above the baseline ASR. However, smaller improvements in validity were obtained with mobile phone recordings. The authors therefore conclude that labour-intensive linguistic-phonetic analysis may be unwarrented in FVC casework. [3] present promising results resolving the false acceptances produced by an i-vector-based ASR using voice quality analysis. The move towards an integrated approach is also highlighted by the inclusion of a human-assisted ASR (HASR) element within the NIST evaluations in 2010 [4]. Further, the use and acceptance of ASRs in conjunction with linguistic-phonetic analysis in casework is increasing, with labs in Germany and Sweden providing conclusions based on combinations of analyses.

In [5] we presented the results of likelihood ratio (LR)-based testing using combinations of different spectral and temporal features extracted from the hesitation markers *uh* and *um*. Hesitation markers are thought to be good speaker discriminants since they occur frequently, are less susceptible to coarticulation than lexical vowels, and display less within-speaker variability since speakers have little conscious control over their production [6,7]. In [5], testing was conducted using single recordings from a set of 60 young male speakers of standard southern British English (SSBE) [8]. Different combinations of input variables for each hesitation type were analysed and compared in terms of strength of evidence and system performance. The best performing system used the F1, F2, and F3 trajectories of the vocalic portion of *um* fitted with quadratic polynomials, together with vowel and nasal durations. This system achieved an equal error rate (EER) of 4.08% and a Log LR cost (C_{llr}) [9] of 0.12. A number of general findings also emerged from these tests. First, *um* consistently performed better than *uh*. Second, the inclusion of information from the first three formants outperformed any individual formant or combination of two formants. Third, modelling the formant trajectories of *um* dynamically (i.e. with multiple measurements across the duration of the vowel) outperformed static midpoint analyses. However, for *uh*, midpoint input outperformed dynamics. Finally, the inclusion of durations consistently improved system performance.

The present study expands on the promising results of [5] to assess the potential additional value of combining MFCC-based ASR systems with the best performing hesitation system, i.e. the formant dynamics and durations of *um*. As in [1] and [2], the ASR acts as a baseline system against which the individual and fused systems are compared. Performance is evaluated in terms of both EER and C_{llr}. This study builds on [5] in a number of ways. The same corpus is used, but two recordings of each speaker in separate forensically relevant tasks are analysed. This provides a more realistic estimation of the within-speaker variability in FVC casework, and therefore a more realistic representation of the performance of the systems under casework conditions. The analysis includes more data per speaker than in [5]. Finally, multiple replications of the same experiment are conducted using randomised sets of speaker.

2. Methodology

2.1. Recordings

Data were drawn from the Dynamic Variability in Speech (DyViS) corpus [8]. DyViS contains male speakers of SSBE aged 18-25. Recordings of Tasks 1 and 2 were used. Task 1 involves a mock police interview in which the participant is questioned about a crime. Task 2 involves an information

exchange task conducted over the telephone between the participant and an accomplice. For this study the high quality, near-end studio recordings of both tasks were used. Both tasks are around 15 minutes in duration. In their design, DyViS tasks 1 and 2 capture the situational differences across recordings (e.g. interlocutor, topic, Lombard speech due to telephone transmission) typical in real FVC casework. The tasks were recorded in separate sessions on the same day. There was thus some time between the two sessions.

2.2. Feature extraction

2.2.1. Linguistic-phonetic system

The linguistic-phonetic system consisted of quadratic polynomial coefficients derived from the F1 to F3 trajectories of the vocalic portion of *um*, as well as vowel and nasal durations. PRAAT TextGrids containing manually segmented tokens of *um* were already available for 88 of the 100 speakers for Task 1. *um* tokens from the Task 2 recordings were also segmented for the same 88 speakers. For both tasks, F1 to F3 values were extracted at +10% steps across each vowel, tracking between five and six formants within a range of 0 to 5kHz. Vowel and nasal durations were also extracted.

The raw data were inspected visually and obvious measurement errors removed. Missing values were replaced with the mean of the values for the adjacent steps. A series of heuristics were then applied to remove less obvious errors. Data points outside specific ranges were removed: 250-900Hz for F1, 900-1900Hz for F2, and 1900-3200Hz for F3. Univariate outliers were calculated based on the group mean at each +10% step. Values of greater than ± 3.29 standard deviations from the mean were removed. Where possible, missing values were again replaced with the mean of adjacent values. Finally, formant trajectories were fitted with quadratic polynomials, generating three coefficients per formant.

Speakers with fewer than 20 tokens per sample were removed, leaving a data set of 63 speakers with between 20 and 49 tokens per sample (mean=38). Although the number of tokens per speaker may appear unrealistically large relative to real case data, the availability of large amounts of data is increasingly common in FVC casework, especially in high profile cases conducted over many months or years.

2.2.2. Automatic system

A generic MFCC-based Gaussian Mixture Model-Universal Background Model (GMM-UBM) system [10] was used as a baseline against which to assess the performance of the *um* and fused systems. Pre-processing was conducted to isolate the speech-active portion of each sample. Recordings were edited manually to remove overlapping speech, interlocutor speech, clicks and background noise. Automatic clipping detection was then run, and clipped sections removed. Finally, voice activity detection was performed using the `voicebox` toolkit in MATLAB to remove silences greater than 100ms. Utterances were then concatenated into a single sample.

The audio were resampled at 10kHz (frequency range = 0-5000Hz) and MFCCs were extracted using the `rastamat` toolkit in MATLAB. A pre-emphasis filter (coefficient value = 0.97) was applied to each sample. Samples were then divided into a series of frames using a 20ms hamming window shifted at 10ms across the duration of the sample, i.e. with 50% overlap between adjacent frames. A Mel filterbank consisting of triangular filters was applied to the power spectrum of the signal for each frame. The energy in each filter was summed and logged, and the log filterbank fitted with a discrete cosine transform (DCT). The coefficients from the DCT are MFCCs. From each frame, 16 MFCCs were extracted. 16 delta and 16 delta-delta coefficients were also appended to the feature vector for each frame. Following [11], data from three frames before and after utterance boundaries were removed.

2.3. Likelihood ratio (LR)-based system testing

Likelihood ratios (LRs) was used to evaluate the performance of the individual and fused systems. The LR is expressed as:

$$LR = \frac{p(E \mid H_p)}{p(E \mid H_d)}, \quad (1)$$

where p is probability, E is evidence, H_p is the prosecution proposition and H_d is the defence proposition. The numerator of the LR is equivalent to the similarity between the suspect and offender samples, while the denominator is equivalent to the typicality (or distinctiveness) of the offender sample relative to patterns in the relevant population [12].

2.3.1. Feature-to-score stage

From the 63 available speakers, 60 were identified and randomly divided into sets of 20 training speakers, 20 test speakers, and 20 reference speakers. Same- (SS; 20) and different-speaker (DS; 190) comparisons were conducted using the training and test sets separately, with the reference set used to calculate typicality. Each comparison generates a LR-like score. Scores for the *um* system were computed using a MATLAB implementation [13] of Aitken and Lucy's multivariate kernel density (MVKD) formula [14]. MVKD models the suspect data with a normal distribution and the reference data with kernel density made up of equally weighted Gaussians for each reference speaker

GMM-UBM scores for the MFCC system were computed using the MSR Identity Toolbox [15]. A 512 Gaussian UBM was trained on data from the 20 reference speakers. Suspect samples for each development and test speaker were created using maximum a posteriori (MAP) adaptation. The suspect data were first modelled as a 512 Gaussian GMM. The GMM is parameterised using the means, variances and weights of the Gaussians. For each suspect, a copy of the UBM is made and then adapted towards the means, variances and weights from the suspect data. This is then used as the suspect model. The score (s) for each suspect-offender comparison is then:

$$s = \frac{1}{T}\sum_{i=1}^{T} \log(p(x_i \mid \lambda_{sus}) - p(x_i \mid \lambda_{bkg})) \quad (2)$$

where T is the number of observations in the offender data, x_i is the offender value, λ_{sus} is the suspect model and λ_{bkg} is the background (reference) model.

2.3.2. Score-to-LR stage

The *um* and MFCC systems were initially analysed separately. For each system, calibration coefficients were calculated from the scores for the training data using logistic regression. The calibration coefficients were then applied to the scores for the test data to produce sets of calibrated log LRs (LLRs). The systems were also combined using logistic-regression fusion. In all cases, calibration and fusion coefficients were calculated using a robust MATLAB implementation [16] of scripts from Brümmer's FOCAL toolkit [17].

2.3.3. System evaluation and replication

The validity of the systems was evaluated using EER and C_{llr} [9]. EER represents the threshold-independent point at which the percentage of false hits (DS providing SS evidence) and misses (SS providing DS evidence) is equal. In this way EER is based on categorical, accept-reject decisions. C_{llr} is a cost function which penalises the system for the magnitude of contrary-to-fact LLRs, such that high magnitude contrary-to-fact LLRs are penalised more heavily than contrary-to-fact LLRs around threshold. The closer the C_{llr} to zero, the better the validity of the system. Testing was repeated using quasi Monte Carlo simulations. 20 different randomised sets of training, test, and reference data were created, and patterns compared across replications.

3. Results

3.1.1. Individual systems

Table 1 displays the mean and range of validity values for the MFCC and *um* systems across the 20 replications. In 17 of the 20 replications the ASR system outperformed the linguistic-phonetic system. The ASR systems produced a mean EER of 2.57% compared with 4.83% for the *um* systems, and a mean C_{llr} of 0.144 compared with 0.261 for the *um* systems.

Table 1. *Mean and range (max-min) of C_{llr} and EER (%) values for the um and MFCC systems across 20 replications.*

System	C_{llr} *Mean*	C_{llr} *Range*	EER *Mean*	EER *Range*
um	0.261	0.751	4.83	8.68
MFCC	0.144	0.526	2.57	5.13

Nonetheless, the results for *um* are extremely promising. First, *um* outperformed the ASR system in three replications, despite the ASR system using information from the entire speech-active portion of the sample. Second, two of the *um* systems outperformed the system in [5] (where EER=4.08% and C_{llr}=0.12) despite the use of separate suspect and offender samples. The remaining 18 replications produced validity very close to that produced in [5]. This suggests that *um* is relatively robust against the type of stylistic variation commonly found across in FVC casework.

For both the linguistic-phonetic and ASR systems, the variability in validity as a function of the configuration of speakers in the training, test, and reference sets is relatively large. At least for C_{llr}, this is, in part, due to two replications which provided atypically poor validity relative to the other replications. However, even excluding these replications the range of validity values is large. The implications of this are discussed in 4.

3.1.2. Fused systems

Figures 1 (C_{llr}) and 2 (EER) display the validity of each of the baseline ASR and fused systems, indicating the direction and magnitude of the change in performance with the addition of the *um* system. The C_{llr} of the fused systems was found to be consistently lower across the 20 replications. The absolute improvement in C_{llr} ranged from 0.003 to 0.43, with mean improvement of 0.09. In terms of percentage improvement, the addition of *um* reduced C_{llr} by between 8.7% and 89.9% relative to the baseline ASR systems. The largest improvement in performance was found for the ASR systems with inherently high C_{llr}. For replication 15, the fusion with the *um* system reduced C_{llr} from 0.55 to 0.12. For the ASR systems with inherently better C_{llr} (i.e. closer to 0), the magnitude of the improvement in the fused system was predictably smaller.

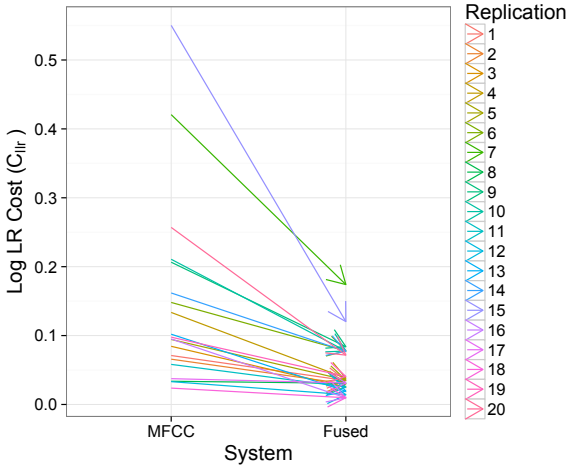

Figure 1: *C_{llr} values for the MFCC-only and fused systems across all 20 replication.*

Similar patterns were found for EER. With the exception of the three ASR systems which produced 0% EER, the remaining 17 fused systems produced lower EER than the baseline system. The absolute improvement in EER ranged from 0.26% to 5.13% (in this replication bringing EER for the fused system down to 0%), with mean improvement of 2.58%.

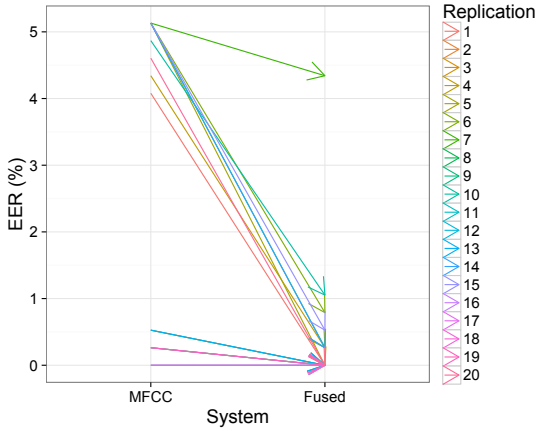

Figure 2: *EER (%) values for the MFCC-only and fused systems across all 20 replication.*

4. Discussion

4.1.1. ASR vs. linguistic-phonetic systems

The individual ASR and linguistic-phonetic systems performed extremely well across the tests conducted in this study, producing mean EER values of less than 5% and mean C_{llr} values of less than 0.27. In 17 of the 20 replications, the ASR system outperformed the linguistic-phonetic system, with the ASR systems optimally achieving an EER of 0% and a C_{llr} of 0.02. The performance of the *um* systems was also very good, optimally achieving an EER of 0.26% and a C_{llr} of 0.08. The extent to which the ASR systems outperformed the

linguistic-phonetic systems is not as great as may be expected, given the considerably larger portion of the recording analysed using the ASR. The results for *um* in 3.1.1. compare very well with previous studies which have considered the performance of formant trajectory-based linguistic-phonetic systems [18,19]. Together with [5], the results offer further support for the value of filled pauses as features in FVC cases. However, somewhat poorer validity for all systems is expected when using more forensically realistic materials, incorporating a greater degree of non-contemporaneity and channel mismatch.

As shown in Table 1, however, for both forms of input the range of variability in system validity across replications is relatively large. This is purely random variation as a function of the particular speakers that make up the training, test, and reference data sets. The speakers themselves all performed the same tasks in the same way and are demographically well matched. The variability across replications is an important issue for FVC evidence, as it may have a considerable effect on the validity of the system presented to the court and the resulting strength of evidence. In the interests of transparency and objectivity, it may be necessary to perform similar replications to assess the sensitivity of system output in real FVC casework. It may then be possible for the expert to present a range of potential system validity values to the court (in the form of a credible interval).

4.1.2. Individual vs. fused systems

Despite *um* producing poorer system validity than the baseline ASRs, very promising improvement was found when the two systems were fused. Improvements in C_{llr} were found across all 20 replications. The mean absolute improvement in C_{llr} was 0.09, equivalent to a mean decrease in C_{llr} of 58.1% relative to the baseline system. Maximally, the fusion of the two systems reduced C_{llr} by 89.9%. Such improvement is considerably greater than that reported in [1]. Improvements in EER were found for 17 of the 20 replications. The three exceptions were the baseline ASR systems which were already performing at ceiling for EER (i.e. they produced 0% EER individually and 0% EER when fused with *um*).

This suggests that the speaker-discriminatory information encoded within the formant dynamics and durations of *um* may be orthogonal to that encoded within the MFCCs and derivatives. Further, the combined systems benefit from the fact that, as well as the input data being potentially independent, the speaker-discriminatory power of both systems independently was very good. This leads to almost complete separation of SS and DS pairs when fused. These results highlight the potential value of informed linguistic-phonetic analysis in FVC, and the importance of considering multiple variables (of different types) in any analysis. However, based on [2], the magnitude of the improvement in such fused systems over baseline ASRs may be less when using more forensically realistic data.

5. Conclusions

This study has shown that the performance of an MFCC-based FVC system can be improved, in some cases considerably, by incorporating the formant trajectories and durations of the vocalic portion of the hesitation marker *um*. These results highlight the value of informed linguistic-phonetic analysis in FVC, and support the move towards integrating the best elements of different methods in order to improve the validity and reliability of FVC evidence presented to courts. Future work will consider the additional benefit of linguistic-phonetic analysis to more state-of-the-art, i-vector ASR systems.

6. Acknowledgements

This research was partly funded by an ESRC PhD Scholarship to Vincent Hughes, a University of York Annual Fund Masters bursary to Sophie Wood, and by the AHRC grant *Voice and Identity* (AH/ M003396/1).

7. References

[1] Zhang, C. and Enzinger, E., "Fusion of multiple formant-trajectory- and fundamental-frequency-based forensic-voice-comparison systems: Chinese /ei1/, /ai2/, and /iau1/", Proc. ICA - POMA 19, 2013.

[2] Zhang, C. et al., "Effects of telephone transmission on the performance of formant trajectory-based forensic voice comparison - female voices", Speech Communication 55(6), 796-813, 2013.

[3] Gonzalez-Rodriguez, J. et al., "What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trial, Proc. Odyssey, pp. 34-40, 2014.

[4] Greenberg, G. et al., "Human assisted speaker recognition (HASR) in NIST SRE2010", Proc. Odyssey, 180-185, 2010.

[5] Anonymous, "Strength of forensic voice comparison evidence from the acoustics of filled pauses", IJSLL 23(1):99-132, 2016.

[6] Tschapse, N., Trouvain, J., Bauer, D. and Jessen, M., "Idiosyncratic patterns of filled pauses", IAFPA Conference, Marrakesh, Morocco, 2005.

[7] Jessen, M., "Forensic phonetics", *Lang. Ling. Compass* 2(4):671-711.

[8] Nolan, F., McDougall, K., de Jong, G. and Hudson, T., "The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research", IJSLL 16(1):31-57, 2009.

[9] Brümmer, N. and du Preez, J., "Application-independent evaluation of speaker detection", Computer Speech and Language 20(2-3):230-275, 2006.

[10] Reynolds, D. et al., "Speaker verification using adapted Gaussian mixture models", Digital Signal Processing 10:19-41, 2000.

[11] Enzinger, E., Morrison, G.S. and Ochoa, F, "A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case", Sci. Jus., 56(1):42-57, 2016.

[12] Aitken, C.G.G. and F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists (2nd ed), Wiley, 2004.

[13] Morrison, G.S., "MATLAB implementation of Aitken and Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation, 2007, Online: http://geoff-morrison.net/#MVKD.

[14] Aitken, C.G.G. and Lucy, D., "Evaluation of trace evidence in the form of multivariate data", Applied Statistics 54:109-122, 2004.

[15] Sadjadi, S.O., Slaney, M. and Heck, L., "MSR Identity toolbox v1.0: a MATLAB toolbox for speaker-recognition research", IEEE, 2013.

[16] Morrison, G.S., "Robust version of train_llr_fusion.m from Niko Brümmer's FOCAL Toolkit", 2009, Online: http://geoff-morrison.net/#TrainFus.

[17] Brümmer, N., " The FOCAL Toolkit", Online: http://niko.brummer.googlepages.com/

[18] Rose, P., "Bernard's 18 – vowel inventory size and strength of forensic voice comparison evidence", Proc. SST, pp. 30-33, 2010.

[19] Rose, P., "Forensic voice comparison with monophthongal formant trajectories – a likelihood ratio-based discrimination of 'schwa' vowel acoustics in a close social group of young Australian females", Proc. ICASSP, pp. 4819-4823, 2015.