Do Long-Term Acoustic-Phonetic Features and Mel-frequency Cepstral Coefficients Provide Complementary Speaker-Specific Information for Forensic Voice Comparison?

Ricky K. W. Chan[1] and Bruce X. Wang[2]

[1]*Speech, Language and Cognition Laboratory, School of English, University of Hong Kong*
[2]*Department of English and Communication, Hong Kong Polytechnic University*
rickykwc@hku.hk, brucex.wang@polyu.edu.hk

Acknowledgements

This study was generously supported by the Hong Kong Research Grants Council Early Career Scheme (HKU Project Code: 21606918). We are also grateful to Prof. Philip Rose for his invaluable feedback at the early stage of this research project, and Dr. Vincent Hughes on his advice on feature fusion.

Correspondence concerning this article should be addressed to Ricky Chan, Speech, Language and Cognition Laboratory, School of English, University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: rickykwc@hku.hk

Do long-term acoustic-phonetic features and mel-frequency cepstral coefficients provide complementary speaker-specific information for forensic voice comparison?

Abstract

A growing number of studies in forensic voice comparison have explored how elements of phonetic analysis and automatic speaker recognition systems may be integrated for optimal speaker discrimination performance. However, few studies have investigated the evidential value of long-term speech features using forensically-relevant speech data. This paper reports an empirical validation study that assesses the evidential strength of the following long-term features: fundamental frequency (F0), formant distributions, laryngeal voice quality, mel-frequency cepstral coefficients (MFCCs), and combinations thereof. Non-contemporaneous recordings with speech style mismatch from 75 male Australian English speakers were analyzed. Results show that 1) MFCCs outperform long-term acoustic phonetic features; 2) source and filter features do not provide considerably complementary speaker-specific information; and 3) the addition of long-term phonetic features to an MFCCs-based system does not lead to meaningful improvement in system performance. Implications for the complementarity of phonetic analysis and automatic speaker recognition systems are discussed.

Keywords

forensic voice comparison, long-term acoustic-phonetic features**,** mel-frequency cepstral coefficients, speech style mismatch, non-contemporaneous recordings, likelihood-ratio

Highlights

• The evidential strength of long-term acoustic phonetic features, MFCCs, and combinations thereof was evaluated.

• Non-contemporaneous recordings with speech style mismatch from 75 male Australian English speakers were analyzed.

• MFCCs consistently outperformed long-term acoustic phonetic features.

• Source and filter features do not provide considerably complementary speaker-specific information.

• The addition of long-term phonetic features to an MFCCs-based system does not lead to meaningful improvement in system performance.

1. Introduction

Forensic voice comparison (FVC) typically involves comparing voices on two recordings: an unknown offender's voice of, for example, hoax calls, threatening messages, phone scams, or conversation with an accomplice, and a known suspect's voice (e.g. from a police interview). The main goal of FVC is to assist investigative bodies (e.g. police) or triers-of-fact (e.g. jury/judge) in deciding if the known and unknown voices belong to the same speaker or different speakers. With advances in speech communication and recording technology, recorded voices are increasingly presented as evidence in court cases worldwide (French & Stevens, 2013).

A primary goal in FVC research is to identify features that are useful for distinguishing speakers in forensic conditions. Most previous studies focused on the speaker-discriminatory performance of individual speech features. However, individual features alone are rarely sufficient for distinguishing speakers. Maximal speaker-discriminatory power should lie in a combination of variables, and forensic practitioners typically analyse a range of features in casework. Research is needed to identify the optimal combination of speaker-specific variables that will allow forensic practitioners to differentiate speakers. To this end, this paper reports an empirical validation study that assesses the evidential strength of the following long-term features: fundamental frequency (F0), formant distributions, laryngeal voice quality, mel-frequency cepstral coefficients (MFCCs), and combinations thereof.

International surveys of FVC practices among law-enforcement agencies and forensic practitioners worldwide (Gold & French, 2011, 2019; Morrison et al., 2016) have reported four broad approaches to data extraction and analysis: auditory phonetic analysis, qualitative spectrographic analysis (i.e. the voiceprint technique), quantitative acoustic-phonetic analysis and analysis by an automatic speaker recognition (ASR) system (with or without human supervision). In practice, forensic practitioners often adopt a combination of these approaches (e.g. auditory-spectrographic, auditory-acoustic-phonetic or semi-automatic; see Morrison & Enzinger (2019) for a discussion). These surveys reveal that the combination of auditory and acoustic-phonetic analysis is the most widely adopted approach to FVC worldwide, whereas human-supervised ASR has become increasingly popular in the past decade.

Auditory-acoustic-phonetic approaches exploit the componentiality of speech in terms of phonetic features such as vowels, consonants, pitch and voice quality (French & Stevens, 2013). These features can often be related to articulatory settings, individual speech habits and/or socially conditioned linguistic behaviours, and each has auditory/acoustic parameter(s)

that can be measured and analysed separately (Jessen, 2021a). On the other hand, ASR in the FVC context generally refers to the use of computer programs to compare voice samples provided to the system, with varying degrees of human intervention (see Hansen & Hasan (2015) for an overview). Typically, ASR systems do not explicitly examine acoustic-phonetic features but make spectral measurements at regular intervals regardless of whether those measurements originate from consonants or vowels. Commonly made measurements in an ASR system are MFCCs, which characterise the shape of the spectrum (see Davis & Mermelstein (1980) for details). MFCC measurements are often made with a fixed window length which is shifted across all the speech materials of the target speaker in the entire recording (Morrison et al., 2018). A number of coefficients (the exact number differs from one system to another) are derived for characterisation of the speech spectrum across frequency. MFCCs may be accompanied by derivative measurements: *deltas*, which capture the rate of change of MFCC values over time, and *double deltas*, the rate of change of delta values over time (Furui, 1986; Morrison et al., 2018), although they are less commonly included in state-of-the-art ASR systems that involve deep neural networks. But the boundary between ASR and acoustic-phonetic approaches is not always clear-cut (e.g. phonetic features such as formant frequencies may be extracted automatically; MFCC measurements may be extracted within particular phonemes) (Morrison & Enzinger, 2019).

The best way to assess the evidential strength of speech features, or the validity of FVC systems[1] in general, lies in empirical validation. There is now considerable regulatory and judicial pressure—e.g. Daubert ruling (1993), the England & Wales Criminal Practice Directions 19 A (CPD, 2015) and UK Crown Prosecution Service (CPS, 2019)—on experts to empirically validate methods used in forensic comparison using data that reflects case conditions. The importance of empirical validation was also addressed and reiterated in reports by key leading forensic institutions such as European Network of Forensic Science Institutes (Willis et al., 2015; Wagner et al., 2022) as well as government bodies, e.g., the US National Academy of Sciences report (National Research Council, 2009) and President's Council of Advisors on Science and Technology (Lander & PCAST Working Group, 2016). Validation in FVC may be conducted for specific legal cases (case-by-case validation). Nonetheless, due to time and resource constraints, it is more efficient for FVC practitioners to

[1] Here an FVC 'system' is broadly defined as a set of procedures that is employed to compare known and unknown voice samples (Morrison, 2013), including database selection, the approach and method for data analysis and statistical modelling (if appropriate) and the framework for evaluating forensic evidence.

rely on existing validation reports (i.e. anticipatory validation research) if the relevant population and conditions are sufficiently similar to the case at hand (Morrison et al., 2021).

The past two decades have witnessed an increasing number of research papers exploring the speaker-discriminatory power of individual speech features such as vowel formants (e.g. Enzinger, 2014; Hughes, 2014; McDougall 2004; Nolan & Grigoras, 2005; Rose, 2007), laryngeal voice quality (e.g. Chan, 2023; Hughes et al., 2019), lexical tones (e.g. Chan, 2016; 2020; accepted; Chan & Wang, 2024; Pingjai, 2019; Rose & Wang, 2016), and F0 (Hudson et al., 2007; Jessen et al., 2005; Kinoshita et al., 2009). In actual forensic casework, however, forensic practitioners rarely rely on the analysis of one single phonetic feature or one single analytical approach. French and Stevens [18] argue that auditory-acoustic-phonetic and human-supervised ASR approaches may be complementary. To this end, recent research has explored how phonetic features and MFCCs may be combined for optimal speaker-discriminatory performance and the extent to which these speech features carry complementary speaker-specific information. Understanding the relationship and potential correlations among speech features is vital for forensic analysts to avoid under- or overestimation of the strength of evidence.

A few studies found that, when an MFCC-based system is used as the baseline, adding acoustic-phonetic-based systems does not considerably improve system validity (e.g. Enzinger et al., 2012; Zhang et al., 2013; Zhang & Enzinger, 2013). For example, using the /iau/ token produced by 60 female speakers of Northeastern Mandarin, Zhang et al. (2013) found that fusing a formant trajectory-based system with a baseline MFCC system generally did not lead to meaningful improvement in system performance. Enzinger et al. (2012) found that, with data from the segment /n/, the addition of voice-source features extracted by GLOTTEX software to an MFCC-based system did not lead to considerable improvement. On the other hand, some studies have demonstrated that the analysis of acoustic-phonetic features may be complementary to an MFCC-based system (e.g. Hughes et al., 2019; 2023). For instance, based on contemporaneous speech data of hesitation marker *um* in Southern British English, Hughes et al. (2023) found that the performance of MFCC-based system could be considerably improved by fusing it with dynamic formant information. The addition of dynamic F0, additive noise parameters, relative harmonics parameters and root mean square amplitude led to variable improvements across 20 replications.

The conflicting results may be attributable to a number of methodological differences such as sample size, the language involved and, notably, the specific phonemes selected for analysis. It remains unclear to what extent results based on the analysis of a specific phoneme

may be generalisable. On the other hand, the effectiveness of combining MFCCs with long-term phonetic features such as long-term formant distributions (LTFDs), long-term fundamental frequency (LTF0), and long-term laryngeal voice quality (LTLVQ), has received very little research attention (see, e.g., Hughes et al. (2019) for an exception). The measurements of these features are typically pooled across the vocalic portions of a recording, and thus these feature are less susceptible to variability stemming from the realisation of specific words or sounds and within-speaker occasion-to-occasion differences. Nolan (1983) identifies long-term quality in speech as a vital source of speaker-specific information. For example, Nolan and Grigoras (2005) argue that LTFDs provide crucial insights into various dimensions of a speaker's vocal tract and a speaker's habitual characteristics such as palatalisation and lip rounding. A number of studies have showed promising speaker-discriminatory performance using LTFDs (e.g. French et al., 2015; Gold et al., 2013; Moos, 2010). On the other hand, LTF0 captures global F0 characteristics of a speaker in a recording, but LTF0 generally show poor evidential strength (e.g. Kinoshita, 2005; Rose & Zhang, 2018), mainly due to high within-speaker variability of F0 as a result of factors such as emotions, state of health, time of recording, and Lombard effect (Braun, 1995). A few recent studies have evaluated the evidential strength of LTLVQ (also known as phonation types; e.g. Chan, 2023; Hughes et al., 2019; Jessen et al., 2023), but in general the speaker-discriminatory performance of spectral tilt parameters and additive noise parameters did not appear to be promising, especially when speech style mismatch and non-contemporaneous recordings were involved. Still, it should be noted that even when individual features show little evidential strength, combining them with other features may still improve overall system performance (Hughes et al., 2023). It is worth exploring whether these long-term phonetic features may be combined with MFCCs for better speaker discrimination.

Phonetic theories may provide insights into the complementarity among MFCCs and the long-term acoustic-phonetic features discussed above. For example, according to the psychoacoustic model of voice quality proposed by Kreiman et al. (2014), harmonic source spectral shape, inharmonic source excitation, time-varying source characteristics, and vocal tract transfer function are separate components that are both necessary and sufficient for modelling perceived voice quality based on empirical findings. This suggests that these components have different characteristics for distinguishing voices. Most of the LTLVQ parameters tested in Chan (2023), Hughes et al. (2019), Jessen et al. (2023) and in the present study correspond to the harmonic source spectral shape (spectral tilt parameters) and the inharmonic source excitation (additive noise parameters) components, whereas LTF0 and

LTFDs are relevant to time-varying source characteristics and vocal tract transfer function respectively. Also, with reference to the source-filter theory (Fant, 1960), LTF0 and LTLVQ can be categorized as 'source' features and LTFDs as filter features, and these two types of features are assumed to be largely independent of each other in speech production (although some evidence of interrelationships between source and filter features was noted by Hughes et al. (2023)). We thus hypothesize that LTLVQ, LTF0, and LTFDs will provide different and considerable complementary information for speaker discrimination. On the other hand, MFCCs are often assumed to mostly capture vocal tract filter information, and it has been asserted that source information is removed by smoothing out rapid local changes in the spectrum that are caused by harmonics or noise in the signal (Hughes et al., 2023; Jurafsky & Martin, 2009). Nonetheless, the degree of source-filter decoupling in MFCCs hinges on the number of coefficients involved in the analysis: smaller numbers of cepstral coefficients lead to a smoother spectral representation, which results in less source information being captured (Hughes et al., 2023). With the use of 13 coefficients in the present study, our MFCC data are expected to carry both source and filter information. Therefore, we hypothesize that the addition of LTF0, LTLVQ and/or LTFDs will not considerably improve MFCCs-based system performance.

Moreover, not all existing studies on speaker-discriminatory performance of speech features have clear and direct implications for FVC. It has been argued that in order for validation results to be directly relevant for forensic casework, the likelihood-ratio framework, which measures the probability of evidence under the prosecution and under the defence hypotheses (i.e. two speech samples being produced by the same speaker or different speakers), should be used for forensic inference (Morrison et al., 2021; Rose, 2002; Saks & Koehler, 2005). Besides, the validation data should be sufficiently large and representative of a given population, and should be sufficiently reflective of conditions found in casework (Morrison et al., 2021). For instance, the speech styles involved should be relevant to forensic situations, and there is often a mismatch in speech style involved in the unknown and the known voices (e.g. spontaneous conversation with an accomplice vs. police interview). Validation data should ideally allow the testing of the extent to which such mismatch may adversely affect the speaker-discriminatory performance of FVC systems. Also importantly, FVC casework typically involves recordings that are separated by weeks, months or even years, and greater within-speaker variation in speech is assumed for such non-contemporaneous data (Morrison et al., 2012). To ensure that results of a validation exercise are forensically relevant, non-contemporaneous recordings should be used to avoid an underestimation of within-speaker

variability and overly optimistic estimates of a system's validity and reliability. However, existing validation studies that involved non-contemporaneous recordings and speech style mismatch are limited, partly because such forensically-oriented databases are scarce. Whilst worse system performance is normally expected when speech style mismatch and non-contemporaneous recordings are involved, the exact impact of these two factors on individual speech features and combinations thereof requires empirical investigation in a bid to understand how they may perform in actual forensic casework.

In sum, taking into account all these factors, the paper reports a likelihood ratio-based validation study of long-term speech features (fundamental frequency (F0), formants, laryngeal voice quality, mel-frequency cepstral coefficients (MFCCs), and combinations thereof) using non-contemporaneous recordings with speech-style mismatch.

2. Methods

2.1 Data

This study used the same dataset reported in Chan (2023). The speech data were from a forensically-oriented database of 231 male Australian English speakers (Morrison et al., 2015). The speakers were recorded on one to three occasions in accordance with the protocol outlined in Morrison et al. (2012). During each of these sessions, they were required to undertake three distinct speaking tasks: engaging in a casual telephone conversation with a friend or colleague (referred to as CNV), exchanging information over the telephone via fax, and participating in a pseudo-police interview (INT). In instances where participants were recorded on multiple occasions, there was an approximate two-week interval between each recording session. The recordings were saved in a high-quality format with noise and cross-talk manually removed.

This study focused on male speakers as males are more commonly involved in crime than females (Steffensmeier & Allan, 1996). Among the 231 male speakers, only speakers recorded more than one occasion were selected given the importance of analysing non-contemporaneous recordings (Morrison et al., 2021). To further control for speakers' age and regional background as far as possible, 75 speakers aged 18–45, mostly from Sydney and other areas within the state of New South Wales (see Appendix A for details), were selected. To test the effects of non-contemporaneous recordings and speech style mismatch, the CNV and INT tasks recorded in two separate sessions (i.e. CNV1, CNV2, INT1, INT2) were analyzed. The

speech styles involved in these two tasks are commonly found in forensic casework (Morrison et al., 2012).

2.2 Feature extraction and parameterization

Vocalic portions of the recordings were manually segmented and labelled in Praat (Boersma, Weenink, 2022), resulting in approximately 33 seconds of net vocalic material per speaker per recording. Acoustic-phonetic features were extracted using VoiceSauce (Shue et al., 2011) and MFCCs were derived using the *librosa* Python library (McFee et al., 2015), with a 20 ms window length and 10 ms window shift. Details are as follows:

1) Long-term fundamental frequency (LTF0): fundamental frequency was extracted using the Straight algorithm (see [36] for technical discussion).

2) Long-term laryngeal voice quality (LTLVQ): We focused on five spectral tilt measures (H1-H2, H2-H4, H1-A1, H1-A2, H1-A3, with harmonic/spectral amplitudes corrected for formant frequencies and bandwidths) and five additive noise measures (cepstral peak prominence (CPP) and harmonic-to-noise ratio (HNR) at four frequency ranges: 0–500 Hz, 0–1500 Hz, 0–2500 Hz, and 0–3500 Hz). See Chan (2023) and Hughes et al. (2019) for discussion on these laryngeal voice quality acoustic measures. We used the same LTLVQ data reported in Chan (2023), but the speaker configurations of the training, test, and reference sets were different from those of Chan (2023) as the 75 speakers were randomly assigned to one of these sets for statistical modelling in each repetition (see Section 2.3 for details).

3) Long-term formant distributions (LTFDs): the first three formants (F1-F3) were extracted using the algorithm in the Snack Sound Toolkit (Sjölander, 2004), with a 6000 Hz ceiling for four formants and a pre-emphasis of 0.96 and 12 LPC order.

4) Mel-frequency cepstral coefficients (MFCCs): the first 13 MFCCs, alongside their corresponding delta and delta-delta coefficients (39 coefficients in total), were derived with a frequency range from 0 to 11,025 Hz.

2.3 Statistical analysis

Outliers, defined as data points that are three median absolute deviations away from the overall media (Leys et al., 2013), were removed as they were not representative of the speakers' typical long-term speech characteristics. F0 values outside 50–300 Hz, F1 values outside 250–900 Hz, F2 values outside 900–2000 Hz, and F3 values outside 1900–3200 Hz were also removed (Hughes et al., 2023; Hudson et al., 2007). The evidential strength of the above parameters and combinations thereof were assessed under the LR framework.

The 75 speakers were first randomly assigned to the training, test or reference set (i.e. 25 speakers in each set). Same-speaker and different-speaker comparisons were conducted for the training and test speakers. Each comparison will produce a score which quantifies the similarity between the two sets of data, and the typicality of the data based on a model created by the reference set. The Gaussian Mixture Model-Universal Background Model (GMM-UBM) (Reynolds et al., 2000) was used for this process. For each feature, GMMs were fitted with varying number of Gaussians (1, 2, 4, 8, 16, 32, 64) in order to determine the option that would fit the data better (Chan, 2023; Jessen, 2021b). Calibration and the conversion of scores to interpretable LRs were conducted using logistic regression (Brümmer et al., 2007). This involves shifting and scaling the test scores using calibration coefficients learnt from the training scores to enhance their comprehensibility, comparability, and interpretability (Morrison et al., 2013). The procedure above was replicated 30 times with different speakers in the training, test and reference sets to test the system reliability of the parameters (Wang et al., 2019). System validity was evaluated based on the distributions of two commonly used metrics across the 30 replications: log-LR cost (C_{llr}) and equal error rate (EER) (see Morrison (2009) for explanations on these two metrics). For both C_{llr} and EER, values closer to zero imply better performance with fewer and less severe speaker-discriminatory errors. A C_{llr} value of 1 or above implies that the system yields no speaker-discriminatory information.

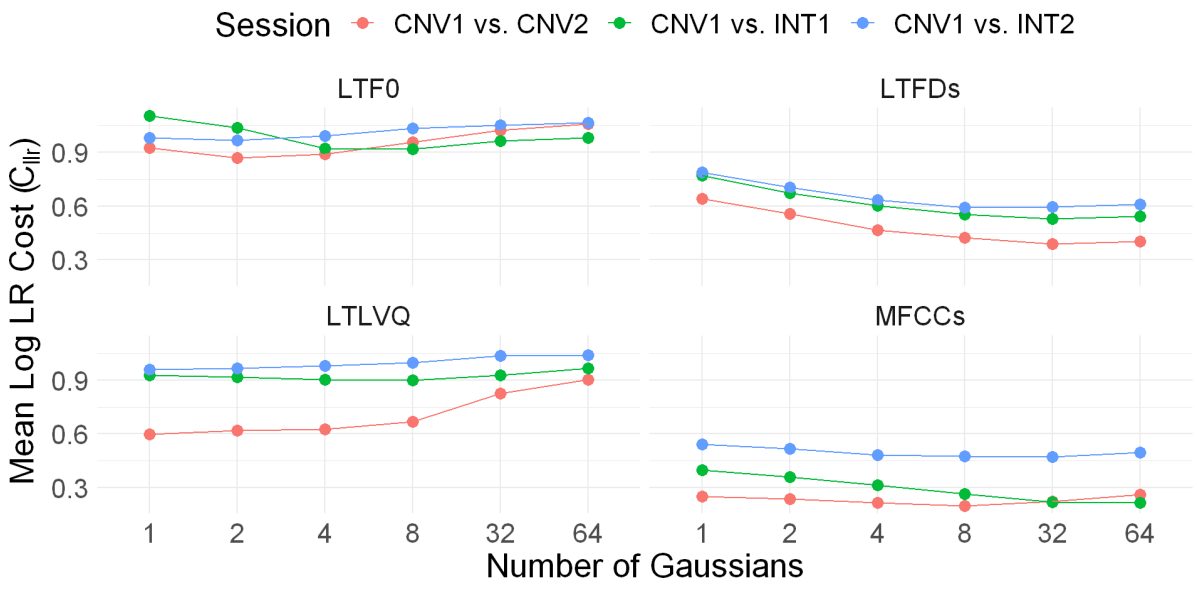
To test if better system performance can be achieved with different combinations of the above features, scores yielded from individual features were fused via logistic regression. To this end, pre-testing was first conducted to identify the optimal number of Gaussians that yielded the best training score for LTF0, LTFDs, LTLVQ and MFCCs separately. Logistic-regression fusion (Brümmer et al., 2007; Pigeon et al., 2000), which takes into account the underlying correlations in these scores, was then applied to these scores to compute the calibrated LRs, C_{llr} and EER for all possible combinations of these features.

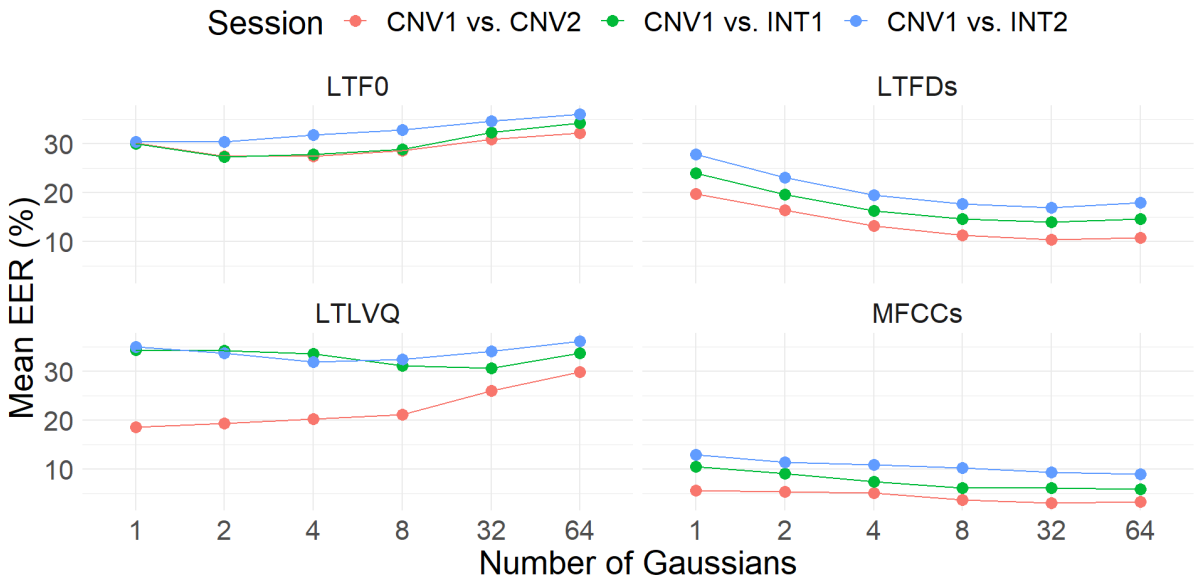
3. Results and discussion

The use of reference population data is curtailed by their limited availability, an often-noted issue among FVC experts (Gold & French, 2019). A recent survey revealed that 68.8 % of FVC practitioners use some form of reference population data, with many reporting that data from published research can be cited in their expert case reports (Gold & French, 2019). Without such data, FVC practitioners must either use self-collected data, which can be time-consuming, labour-intensive and subject to bias, or rely on their intuition or experience, which is far from precise or reliable. Here, mean values and standard deviations of F0, F1, F2 and F3 from the 75 speakers in CNV1, CNV2, INT1, and INT2 can be found in Appendix B. Those of individual acoustic voice quality parameters can be found in Appendix B of Chan (2023). It is hoped that descriptive statistics of these features among Australian English speakers can be useful for forensic practitioners in future casework (e.g. in assessing the typicality of these features as part of their evidence interpretation).

Figures 1 and 2 show the mean C_{llr} and EER values of individual long-term features with varying number of Gaussians in GMM-UBM across 30 repetitions; this serves as a pre-testing for identifying the optimal number of Gaussians for modelling each long-term feature prior to logistic-regression fusion. The number of Gaussians that yielded the lowest C_{llr} values (i.e. better system validity) clearly depends on the features being modelled. Specifically, more Gaussians are required for optimal modelling of LTFDs and MFCCs which mainly capture vocal tract filter information, whereas fewer Gaussians (1−8) seem to be sufficient for LTF0 and LTLVQ which are source features. However, speech style and the time gap between recordings may also play a role and have contributed to the fluctuations observed. Table 1 provides a summary on the optimal number of Gaussians for each feature.

One might be intrigued by the fact that the optimal number of Gaussians required for modelling our MFCCs data appear to be small (8−64). As noted in the Introduction section, MFCC measurements are typically obtained from all speech materials of the target speaker in the entire recording (Morrison et al., 2018). This means that different types of sounds which have different spectral information may need to be modelled with a large number of Gaussians. In the present study, however, MFCC measurements were only made in the vocalic portions of the recordings. This may partly explain why a smaller optimal number of Gaussians were required for our MFCC data.



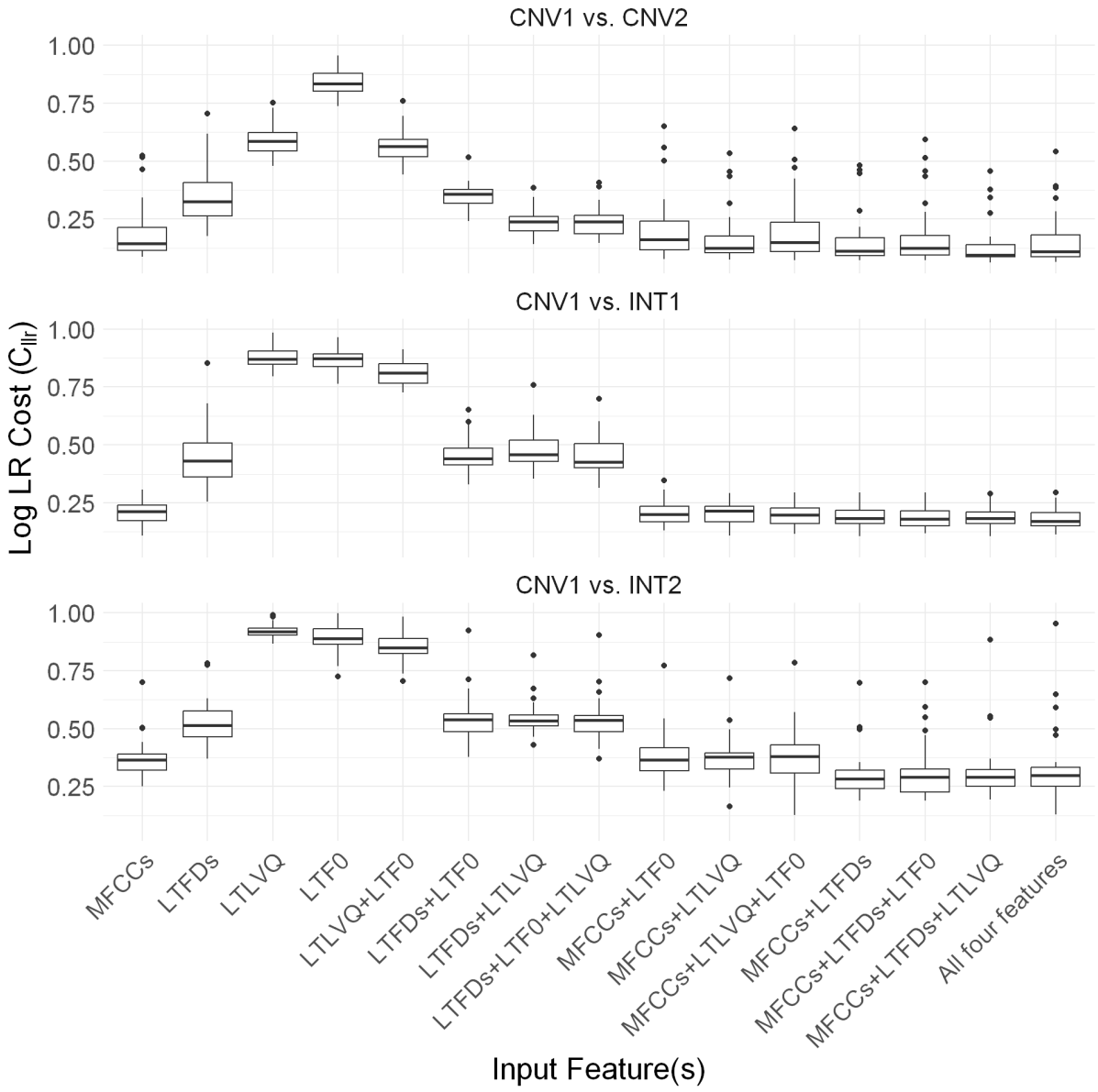


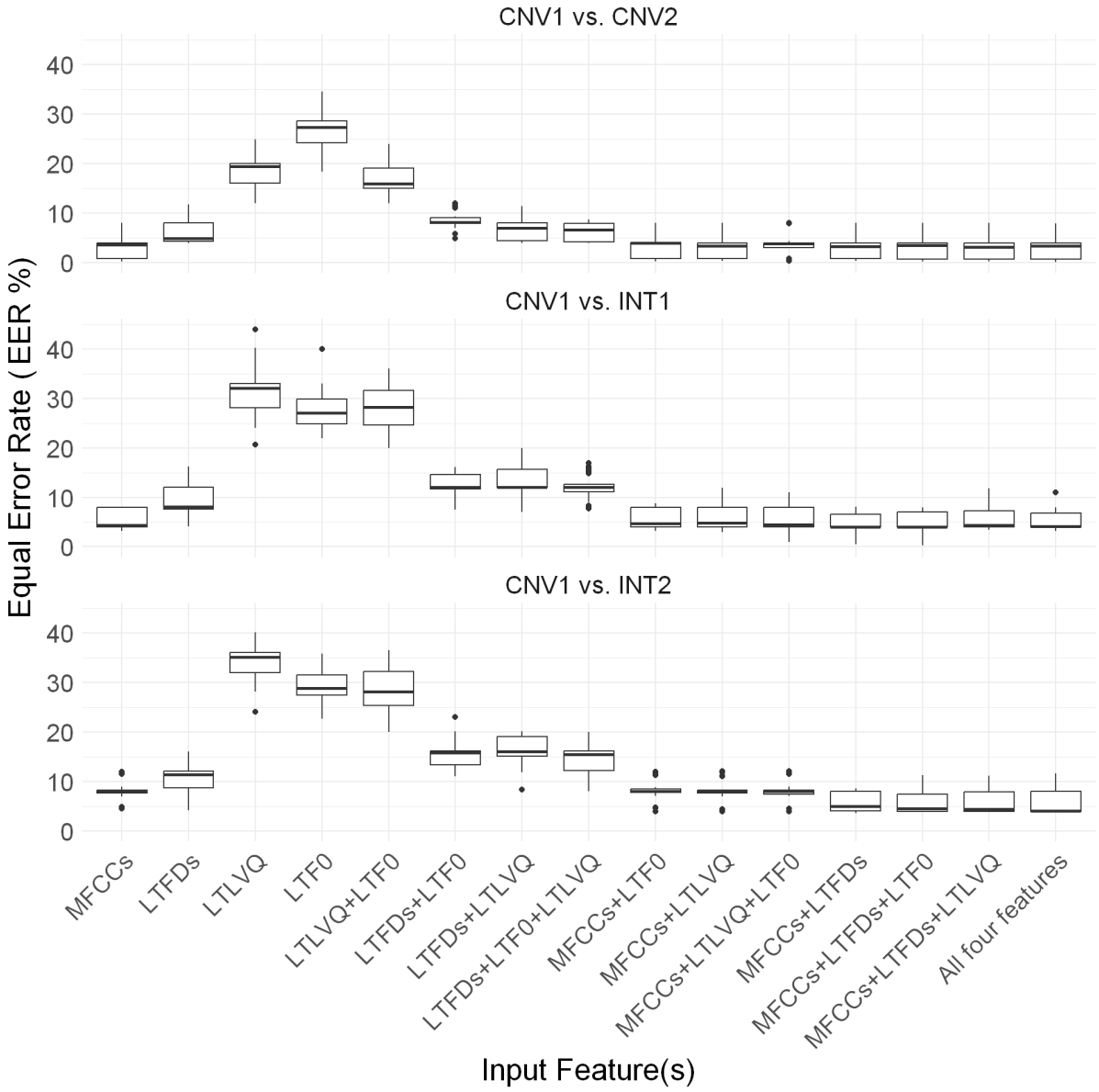
Figures 1 and 2: Mean C_{llr} and EER values of systems based on individual long-term features—LTF0, LTFDs, LTLVQ, and MFCCs—for CNV1 vs. CNV2, CNV1 vs. INT1, and CNV1 vs. INT2.

Session	**Feature & Optimal Number of Gaussians**			
	MFCCs	LTFDS	LTF0	VQ
Conversation 1 vs. Conversation 2	8	32	2	1
Conversation 1 vs. Interview 1	64	32	8	8
Conversation 1 vs. Interview 2	32	32	2	1

Table 1: summary of the optimal number of Gaussians for individual long-term features across conditions

Figures 3 and 4 show the distributions of C_{llr} and EER values respectively for MFCCs, LTF0, LTFDs, LTLVQ, and combinations thereof across 30 repetitions. Tables 2 to 4 provide the descriptive statistics of the corresponding C_{llr} and EER values. The C_{llr} values and EER for MFCCs, LTF0, LTFDs, LTLVQ are based on the number of Gaussians that generated the lowest mean C_{llr} values. For the fusion among the long-term features, the training scores from these best-performing systems were used to train a logistic regression model where the model coefficients were applied to the corresponding test scores to generate calibrated LRs, C_{llr} and EER.





Figures 3 and 4: distributions of C_{llr} and EER values of individual long-term feature and combinations thereof.

CNV1 vs. CNV2

	C_{llr}				EER (%)			
Input Feature(s)	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
MFCCs	0.09	0.52	0.19	0.12	0.17	8.00	3.06	2.22
LTFDs	0.18	0.70	0.35	0.12	4.00	11.75	6.15	2.14
LTLVQ	0.48	0.75	0.59	0.07	12.00	24.92	18.28	3.16
LTF0	0.74	0.95	0.84	0.05	18.42	34.58	26.64	3.79
LTLVQ+LTF0	0.44	0.76	0.57	0.07	11.92	23.92	16.39	3.18
LTFDs+LTF0	0.24	0.52	0.35	0.06	4.92	11.92	8.72	1.77
LTFDs+LTLVQ	0.14	0.38	0.24	0.06	4.00	11.33	6.38	2.19
LTFDs+LTF0+LTLVQ	0.15	0.41	0.24	0.07	4.00	8.75	6.17	1.85
MFCCs+LTF0	0.08	0.65	0.22	0.15	0.17	8.00	3.22	2.26

MFCCs+LTLVQ	0.07	0.53	0.17	0.12	0.33	8.00	3.07	1.98
MFCCs+LTLVQ+LTF0	0.07	0.64	0.20	0.14	0.33	8.00	3.43	2.06
MFCCs+LTFDs	0.07	0.48	0.16	0.11	0.33	8.00	2.69	2.14
MFCCs+LTFDs+LTF0	0.07	0.59	0.19	0.14	0.25	8.00	2.90	2.32
MFCCs+LTFDs+LTLVQ	0.06	0.46	0.14	0.10	0.17	8.00	2.68	2.07
All four features	**0.06**	**0.54**	**0.16**	**0.12**	**0.08**	**7.92**	**3.00**	**2.16**

CNV1 vs. INT1

	C_{llr}				EER (%)			
Input Feature(s)	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
MFCCs	0.11	0.31	0.21	0.05	3.25	8.00	5.63	1.95
LTFDs	0.25	0.85	0.45	0.13	4.08	16.17	9.42	3.16
LTLVQ	0.80	0.99	0.88	0.05	20.67	43.92	31.59	4.47
LTF0	0.76	0.96	0.87	0.04	21.92	39.92	27.74	3.78
LTLVQ+LTF0	0.73	0.91	0.81	0.06	19.92	36.00	27.94	4.52
LTFDs+LTF0	0.33	0.65	0.46	0.08	7.50	16.08	12.41	2.41
LTFDs+LTLVQ	0.35	0.76	0.48	0.09	7.00	20.00	13.22	2.83
LTFDs+LTF0+LTLVQ	0.31	0.70	0.45	0.08	7.75	16.92	12.04	2.61
MFCCs+LTF0	0.13	0.35	0.21	0.05	3.17	8.75	5.63	1.89
MFCCs+LTLVQ	0.11	0.29	0.21	0.05	3.00	11.92	5.85	2.25
MFCCs+LTLVQ+LTF0	0.12	0.30	0.20	0.04	1.00	11.00	5.55	2.29
MFCCs+LTFDs	0.11	0.29	0.19	0.05	0.50	8.08	4.89	1.96
MFCCs+LTFDs+LTF0	0.12	0.30	0.19	0.05	0.33	8.00	4.94	2.06
MFCCs+LTFDs+LTLVQ	0.11	0.29	0.19	0.05	3.42	11.83	5.34	2.08
All four features	**0.11**	**0.30**	**0.18**	**0.05**	**3.17**	**11.00**	**5.16**	**1.98**

CNV1 vs. INT2

	C_{llr}				EER (%)			
Input Feature(s)	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
MFCCs	0.25	0.70	0.37	0.09	4.58	12.00	8.12	1.81
LTFDs	0.37	0.78	0.53	0.10	4.17	16.00	11.06	2.83
LTLVQ	0.87	0.99	0.92	0.03	24.08	40.17	34.09	3.53
LTF0	0.73	1.00	0.89	0.06	22.67	35.83	29.29	3.36
LTLVQ+LTF0	0.71	0.98	0.85	0.06	20.00	36.58	29.05	4.78
LTFDs+LTF0	0.38	0.92	0.54	0.10	11.08	23.00	15.44	2.69
LTFDs+LTLVQ	0.43	0.82	0.55	0.07	8.33	20.08	16.10	3.00
LTFDs+LTF0+LTLVQ	0.37	0.90	0.54	0.10	8.00	20.00	14.78	2.97
MFCCs+LTF0	0.23	0.77	0.38	0.11	4.00	12.00	8.21	2.20
MFCCs+LTLVQ	0.17	0.72	0.37	0.10	4.00	12.08	8.26	2.10
MFCCs+LTLVQ+LTF0	0.13	0.79	0.38	0.12	4.00	12.08	8.00	2.23
MFCCs+LTFDs	0.19	0.70	0.30	0.10	3.67	8.58	5.88	1.86
MFCCs+LTFDs+LTF0	0.19	0.70	0.32	0.12	3.92	11.25	5.70	2.02
MFCCs+LTFDs+LTLVQ	0.20	0.88	0.32	0.13	4.00	11.17	5.56	2.02
All four features	**0.13**	**0.95**	**0.33**	**0.16**	**3.75**	**11.67**	**5.74**	**2.52**

Tables 2-4: descriptive statistics of C_{llr} and EER values for MFCCs, LTF0, LTFDs, LTLVQ, and combinations thereof across 30 repetitions.

Generally speaking, all features and their combinations yielded small standard deviations in C_{llr} and EER values, ranging from 0.03 to 0.16 for C_{llr} and from 1.77 % to 4.78 % for EER, suggesting high system reliability across 30 repetitions with different speaker compositions in the training, test, and reference sets. The results also show clear detrimental effects of speech style mismatch and non-contemporaneous recordings on system validity, and all the features and their combinations display similar patterns. They generally performed best in CNV1 vs. CNV2, followed by CNV1 vs. INT1 and then CNV1 vs. INT2. For the sake of simplicity, the discussion below will focus on C_{llr} values as the key metric on system validity.

As for individual long-term features, LTLVQ performed relatively well when only non-contemporaneous recordings were involved (mean C_{llr} 0.59 in CNV1 vs. CNV2), but considerably worse when there was speech style mismatch (mean C_{llr} 0.88 and 0.92 in CNV1 vs. INT1/INT2 respectively). By contrast, the degree of deterioration in system performance was much smaller for LTF0, LTFDs and MFCCs. Whilst LTF0 may be affected by speech style mismatch and non-contemporaneous recording, its performance was already bad in CNV1 vs. CNV2 (with mean C_{llr} values close to 1) and there was little room for further performance deterioration. LTFDs performed the best among acoustic-phonetic features, with 0.35–0.53 in mean C_{llr}. These results are in line with previous findings that LTFDs, which are supposed to capture vocal tract filter information, are a reasonably good speaker discriminant (e.g. French et al., 2015; Gold et al., 2013; Jessen et al., 2014; Moos, 2010), but source features such as LTLVQ or LTF0 alone carry limited speaker-discriminatory information, especially when speech style mismatch and non-contemporaneous recordings are involved (e.g. Chan, 2023; Jessen et al., 2023; Rose & Zhang, 2018). On the other hand, MFCCs alone returned promising mean C_{llr} values between 0.19 and 0.21 in CNV1 vs. CNV2 and CNV1 vs. INT1 respectively. Their performance dropped to an average of 0.37 in C_{llr} in CNV1 vs. INT2, but still outperformed the other long-term acoustic-phonetic features.

Since forensic practitioners rarely rely on the analysis of a single speech feature in actual casework, we also tested if the four long-term features can be fused in various ways to improve system performance. First, fusing LTF0 and LTLVQ led to slight decreases in mean C_{llr} values across all conditions. This is consistent with the psychoacoustic model of voice quality proposed by Kreiman et al. (2014; 2021) where F0, spectral tilt and additive noise are separate and complementary components of the source characteristics of voice quality.

Second, the addition of LTF0 to LTFDs-based systems brought about no or a very small drop of 0.01 in mean C_{llr} when compared with LTFDs-only systems. The addition of LTLVQ to LTFDs- or LTFDs + LTF0-based systems only improve system validity considerably in CNV1 vs. CNV2 (with a drop in 0.12 in mean C_{llr}) but not in the other two conditions where LTLVQ alone did not perform well. These results suggest that long-term source and filter features do not necessarily provide considerable complementary speaker-discriminatory information, especially when speech style mismatch and non-contemporaneous recordings are involved. Our predictions based on Kreiman et al.'s (2014; 2021) psychoacoustic model of voice quality and the independence between source and filter are not fully supported. Yet, Hughes et al. (2023) reported that the addition of source features to filter-based systems can improve speaker discrimination under optimal conditions, potentially due to the independency between source and filter according to the source-filter theory (Fant, 1960). The conflicting results might be attributed to the fact that, unlike in Hughes et al. (2023), LTF0 and LTLVQ performed rather poorly in our study and probably did not have considerable speaker-discriminatory information to add to LTFDs-based system. Other methodological differences in the two studies should also be noted (the varieties of English involved—southern British English vs. Australian English, the number of speakers, the analysis of hesitation markers vs. the entire recordings, etc.). Future research should explore how source and filter features may be complementary in other types of speech data or conditions (e.g. with other languages, speech styles, channel mismatch and/or background noise).

Lastly, when compared with MFCCs-only systems, the addition of LTF0 and/or LTLVQ to MFCCs-based systems (i.e. MFCCs + LTF0, MFCCs + LTLVQ or MFCCs + LTLVQ + LTF0) generally led to worse or very limited improvements in system validity across conditions. This conflicts with some of the previous findings that the addition of source features may improve speaker discrimination of MFCCs-based systems (e.g. Hughes et al, 2019; 2023). The addition of LTFDs to the systems above (i.e. MFCCs + LTFDs, MFCCs + LTFDs + LTF0, MFCCs + LTFDs+ LTLVQ, and MFCCs + LTFDs+ LTLVQ + LTF0) generally had little or no effect on mean C_{llr} value, consistent with the findings by Becker (2012) and Hughes et al. (2017) that the addition of LTFDs to an MFCC-based ASR system led to very little improvement. Even when this led to a decrease in mean C_{llr}, the drops involved were not greater than 0.07. These results may be partly explained by the fact that long-term MFCCs (13 coefficients and their derivatives) capture information about the overall shape of the spectrum, which overlaps with information captured by LTFDs and LTLVQ such as spectral peaks and

relative harmonics. The extra information from these long-term acoustic-phonetic features did not considerably improve MFCCs-based system performance.

Overall, our analysis of long-term speech features provides no support for the claim that the acoustic-phonetic and ASR approaches may be complementary for speaker discrimination (c.f. French & Stevens, 2013; Hughes et al., 2019; 2023). However, there are other ways where elements of phonetic and ASR approaches may be integrated for improving the task of FVC. Nolan (1983) distinguishes between two mechanisms that shape the speech signal: the linguistic (cognitive mechanism) and the vocal (physical) mechanism. The vocal mechanism consists of the organs and articulators involved in speech production, whereas the linguistic mechanism involves various components of a speaker's linguistic knowledge such as lexicon, syntax, phonetics, phonology, and the set of conventions which the speaker share with the relevant speech community. Whilst speaker-related indexical information is imprinted in both mechanisms, MFCCs and other ASR features that rely heavily on spectral coding predominantly capture information about the features and activities of the vocal mechanism, but much less information on the linguistic mechanism (Nolan, 2022). The analysis of acoustic-phonetic features also strongly reflects the settings and activities of the vocal mechanism (Nolan, 2022). This may explain why the addition of long-term phonetic features did not improve MFCCs-based system performance as long-term acoustic phonetic analysis may at best provide overlapping and corroborative information related to the speakers. However, although long-term speech features may encode rich speaker-specific information (Nolan, 1983), the analysis of long-term features may obscure speaker-idiosyncrasy embedded in the realisation of specific sounds/words. In fact, with the analysis of hesitation marker *um*, Hughes et al. (2023) reported that acoustic-phonetic features can be added to an MFCCs-based system to improve speaker-discriminatory performance. Future studies may explore how phonetic and ASR approaches may be complementary in the analysis of short-term features or speaker-related short-term conditions such as the effects of health and psychological states. On the other hand, auditory phonetic analysis focuses heavily on the linguistic determinants of the speech signal and can reveal indexical information imprinted in the linguistic mechanism such as a speaker's social and dialectal background. Therefore, auditory phonetic analysis should in principle provide partly complementary speaker-specific information about the linguistic mechanism not fully captured by ASR approaches. For example, it has been reported that errors from ASR systems may be resolved through auditory analysis by trained phoneticians, with laryngeal voice quality being a key diagnostic (Hughes et al., 2017; González-Rodríguez et al, 2014). Lastly, given the variability in the speech signal stemming from the plasticity of speech

production and various forensic-relevant conditions, future research should continue to explore how such variability may affect the complementarity of the phonetic and ASR approaches in a bid to fully understand how the two approaches may be best integrated.

4. Conclusions

The paper reports an anticipatory LR-based validation study that empirically evaluates the evidential value of four long-term features—LTF0, LTFDs, LTLVQ, MFCC—and their combinations. The effects of speech style mismatch and non-contemporaneous recordings on system performance were also tested. We found that MFCCs consistently outperformed acoustic-phonetic features across all conditions. Moreover, our results provide no strong evidence that source and filter features in speech necessarily carry complementary speaker-specific information, or that the addition of acoustic-phonetic features to an MFCC-based system would lead to considerable improvement in system validity. Further research should explore how phonetic analysis and ASR systems may be complementary in other ways.

References

- Becker, T. (2012). *Automatischer forensischer Stimmenvergleich*. BoD–Books on Demand.
- Boersma, P. & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.21, retrieved 1 October 2022 from http://www.praat.org/
- Braun, A. (1995). Fundamental frequency: how speaker-specific is it?. *Beiträge zur Phonetik und Linguistik*, *64*, 9-23.
- Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., ... & Strasheim, A. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2072-2084.
- Chan, R. (2016). Speaker variability in the realization of lexical tones. *International Journal of Speech, Language and the Law, 23*(2), 195-214.
- Chan, R. (2020). Speaker discrimination: citation tones vs. coarticulated tones. *Speech Communication*, *117*, 38-50.
- Chan, R. (2023). Evidential value of voice quality acoustics in forensic voice comparison. *Forensic Science International*, *348*, 111725.
- Chan, R. (accepted). Tone languages. In F. Nolan, K. McDougall & T. Hudson (Eds), *Oxford Handbook of Forensic Phonetics*. Oxford University Press.
- Chan, R., Wang, B. (2024). Tone modelling for speaker discrimination. *Language and Speech*. Advance online publication. **https://doi.org/10.1177/00238309241261702**
- CPD. (2015). *England & Wales Criminal Practice Directions*. https://www.justice.gov.uk/courts/procedure-rules/criminal/docs/2015/crim-practice-directions-V-evidence-2015.pdf
- CPS. (2019). *UK Crown Prosecution Service*. https://www.cps.gov.uk/legal-guidance/expert-evidence
- Daubert vs. Merrel Dow Pharms Inc., 1993 509 U.S. 579, 113S. CT 2786.
- Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357 – 366.
- Enzinger, E. (2014). A first attempt at compensating for effects due to recording-condition mismatch in formant-trajectory-based forensic voice comparison. In *Proceedings of the 15th Australasian International Conference on Speech Science and Technology* (pp. 133-136). Australasian Speech Science and Technology Association.

- Enzinger, E., Zhang, C., & Morrison, G. S. (2012). Voice source features for forensic voice comparison-An evaluation of the glottex software package. In *Odyssey 2012 The Speaker and Language Recognition Workshop*.
- Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.
- French, P., Foulkes, P., Harrison, P., Hughes, V., & Stevens, L. (2015). The vocal tract as a biometric: output measures, interrelationships, and efficacy. In *Proceedings of the 18th International Congress of Phonetic Science (ICPhS)*. Glasglow, United Kindgom.
- French, P. & Stevens, S (2013) Forensic speech science. In Jones, M. & Knight, R. (eds.) *The Bloomsbury Companion to Phonetics*. London: Continuum.
- Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *34*, 52 – 59.
- Gold, E., & French, P. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech, Language and the Law*, *18*(2), 293–307.
- Gold, E., & French, P. (2019). International practices in forensic speaker comparisons: second survey. *International Journal of Speech, Language and the Law*, *26*(1), 1–20.
- Gold, E., French, P., & Harrison, P. (2013, June). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. In *Proceedings of Meetings on Acoustics* (Vol. 19, No. 1). AIP Publishing.
- González-Rodríguez, 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. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 33–40.
- Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, *32*(6), 74-99.
- Hudson, T., de Jong, G., McDougall, K. & Nolan, F. (2007). f0 statistics for 100 young male speakers of standard Southern British English. In Trouvain, J. & Barry, W. J. (eds.) In *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrucken, Germany, pp. 1809–1812.
- Hughes, V. (2014). The Definition of the Relevant Population and the Collection of Data for Likelihood Ratio-based Forensic Voice Comparison. Doctoral dissertation, University of York, UK.

- Hughes, V., Cardoso, A., …& Harrison, P. (2019). Forensic voice comparison using long-term acoustic measures of laryngeal voice quality. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*. Melbourne, Australia.
- Hughes, V., Cardoso, A., Foulkes, P., French, P., Gully, A., & Harrison, P. (2023). Speaker-specificity in speech production: The contribution of source and filter. *Journal of Phonetics*, *97*, 101224.
- 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. *Proceedings of Interspeech*, Stockholm, Sweden, 3892–3896.
- Jessen, M. (2021a). Speaker profiling and forensic voice comparison. In Coulthard, M., May, A., & Sousa-Silva, R. (Eds.). *The Routledge Handbook of Forensic Linguistics* (pp. 382-399). New York: Routledge.
- Jessen, M. (2021b). MAP Adaptation Characteristics in Forensic Long-Term Formant Analysis. In *Interspeech*, pp. 411-415.
- Jessen, M., Alexander, A., & Forth, O. (2014, June). Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software. In *Audio Engineering Society Conference: 54th International Conference: Audio Forensics*. Audio Engineering Society.
- Jessen M., Konrat, C., & Horn, J. (2023). Voice comparison analysis of forensic recordings using the VoiceSauce program. In *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*. Prague, Czech Republic.
- Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, *12*, 174–213.
- Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: Computational Linguistics and Speech Recognition* (2nd ed.). New Jersey: Prentice-Hall.
- Kawahara, H., Agiomyrgiannakis, Y., & Zen, H. (2016). Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis. *arXiv preprint arXiv:1605.07809*.
- Kinoshita, Y. (2005). Does Lindley's LR estimation formula work for speech data? Investigation using long-term f0. *International Journal of Speech, Language and the Law*, *12*(2), 235–254.

- Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech, Language and the Law*, *16*, 91–111.
- Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *loquens*, *1*(1), e009.
- Kreiman, J., Lee, Y., Garellek, M., Samlan, R., & Gerratt, B. R. (2021). Validating a psychoacoustic model of voice quality. *The Journal of the Acoustical Society of America*, *149*(1), 457-465.
- Lander, E. S., & PCAST Working Group. (2016). Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, *49*(4), 764-766.
- McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English/aI. *International Journal of Speech, Language and the Law*, *11*(1), 103-130.
- McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, *8*, 18-25.
- Moos, A. (2010) Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician*, *101*: 7-24.
- Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, *49*(4), 298-308.
- Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, *45*(2), 173-197.
- Morrison, G. S., & Enzinger, E. (2019). Introduction to forensic voice comparison. In Katz, W.F., Assmann, P.F. (Eds.) *The Routledge Handbook of Phonetics* (ch.21, pp. 599-634). London: Routledge.
- Morrison, G. S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., ... & Anonymous, B. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, *61*(3), 299-309.

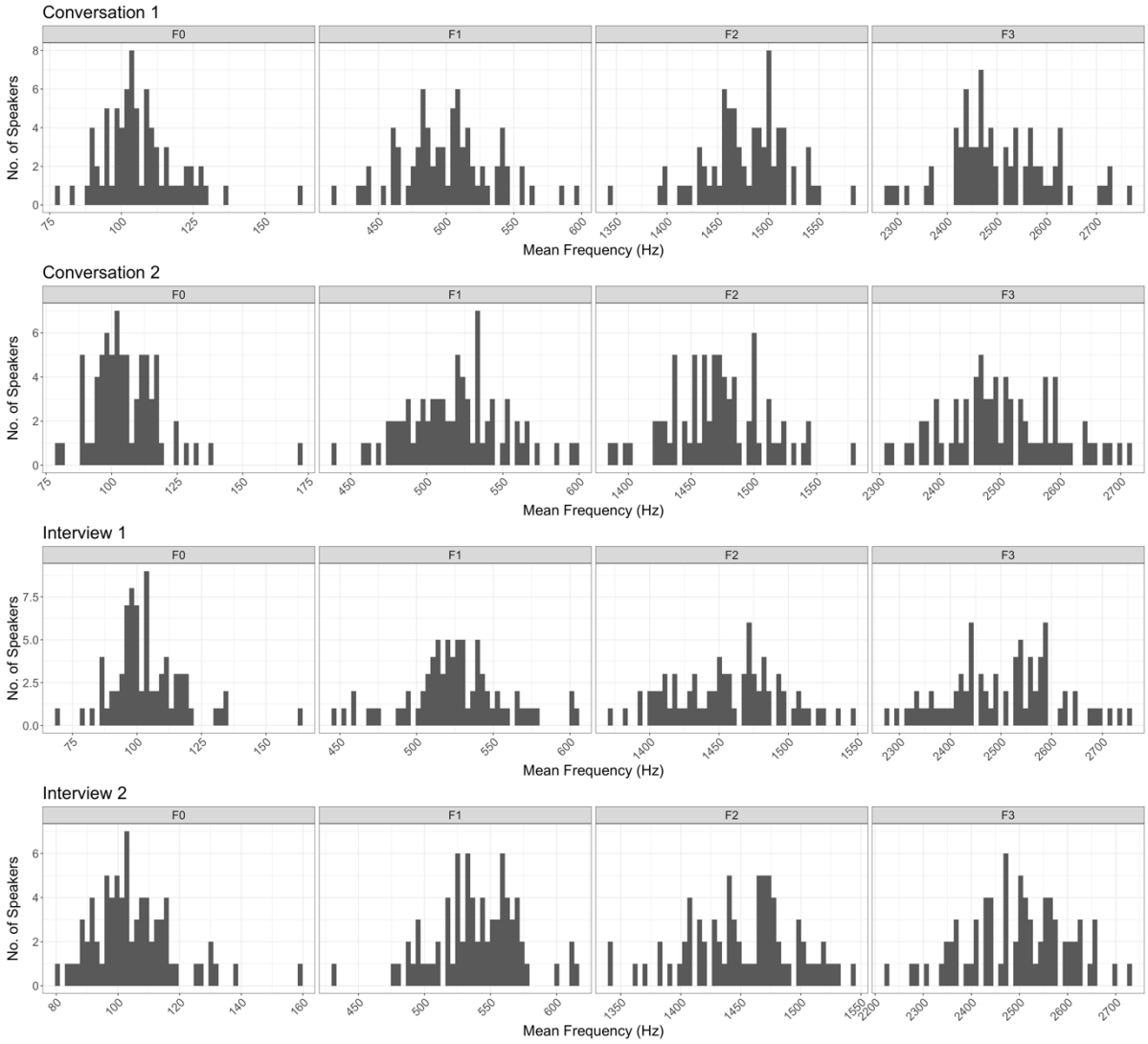
- Morrison G.S., Enzinger E., Zhang C. (2018). Forensic speech science. In Freckelton I., Selby H. (Eds.), *Expert Evidence* (Ch. 99). Sydney, Australia: Thomson Reuters.
- Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Dorny, C. G. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, *263*, 92-100.
- Morrison, G.S., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., Folkes, B. K., De Souza, S., Cummins, N., & Chow, D. (2015). Forensic database of voice recordings of 500+ Australian English speakers.
- National Research Council. (2009). *Strengthening Forensic Science in the United States: A Path Forward*. National Academies Press.
- Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.
- Nolan, F. (2022). Will forensic speech scientists still need ears? [Keynote presentation]. The 30th Annual Conference of the International Association for Forensic Phonetics and Acoustics. Prague, Czech Republic.
- Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, *12*(2), 143-173.
- Pigeon, S., Druyts, P., & Verlinde, P. (2000). Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing*, *10*(1-3), 237-248.
- Pingjai, S. (2019). *A Likelihood-Ratio Based Forensic Voice Comparison in Standard Thai*. PhD Thesis. Australian National University.
- Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, *10*(1-3), 19-41.
- Rose, P. (2002). *Forensic Speaker Identification*. CRC Press.
- Rose, P. (2007). Forensic speaker discrimination with Australian English vowel acoustics. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, pp. 1817–1820.
- Rose, P., & Wang, X. (2016). Cantonese forensic voice comparison with higher level features: likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone. *Odyssey 2016*, 326-333.
- Rose, P., & Zhang, C. (2018). Conversational style mismatch: its effect on the evidential strength of long-term F0 in forensic voice comparison. *Proceedings of ASSTA*, 157-160.

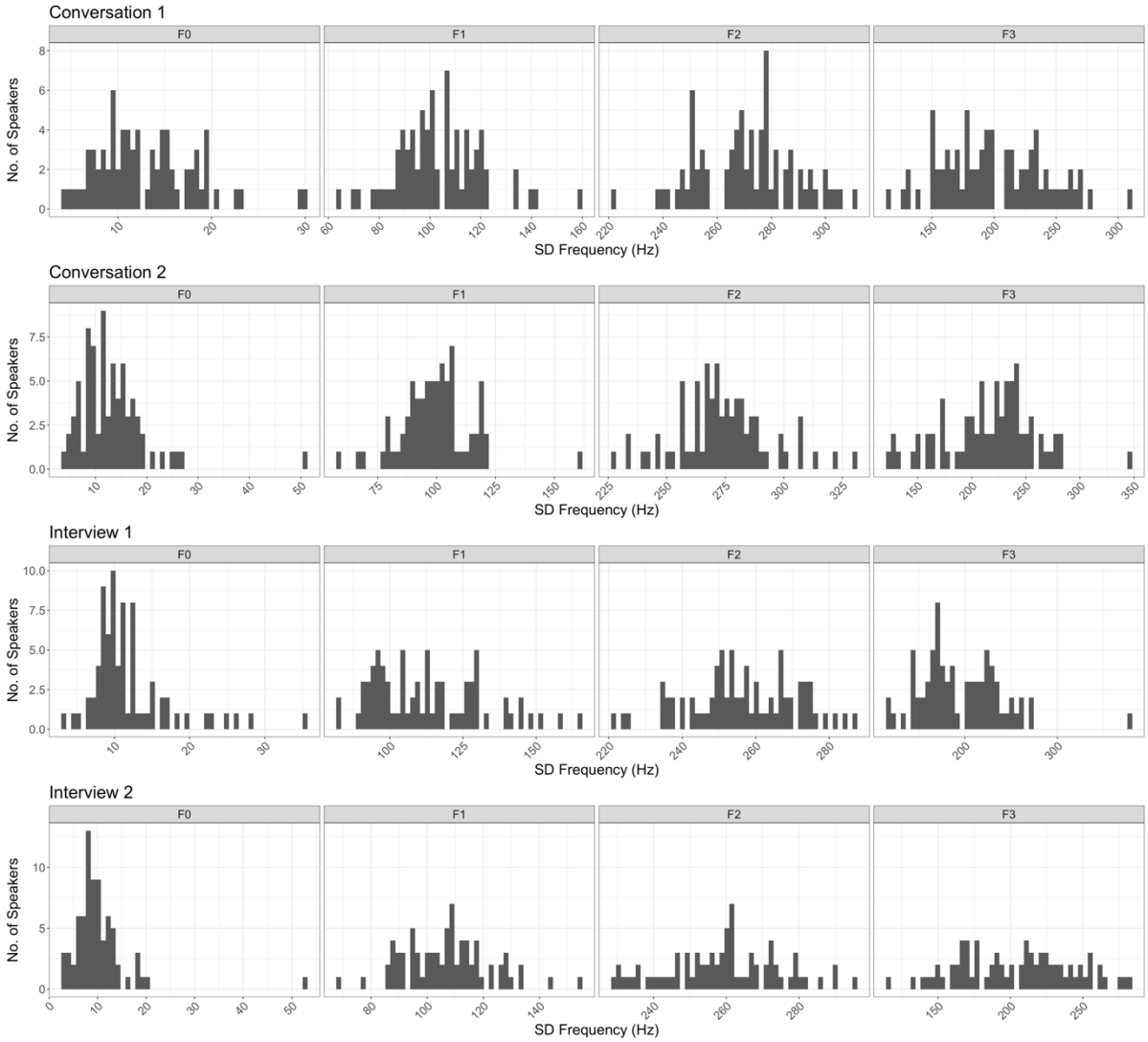
- Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892-895.
- Shue, Y.-L., P. Keating , C. Vicenik, K. Yu (2011) VoiceSauce: A program for voice analysis, *Proceedings of the 17th International Congress of Phonetic Sciences*, 1846-1849.
- Sjölander, K. (2004). The snack sound toolkit [computer program].
- Steffensmeier, D., & Allan, E. (1996). Gender and crime: Toward a gendered theory of female offending. *Annual Review of Sociology*, *22*(1), 459-487.
- Wagner, I., Boss, D., Hughes, V., Svirava, T., Siparov, I., & Rolfes, M. (2022). *Best Practice Manual for the Methodology of Forensic Speaker Comparison*. https://enfsi.eu/wp-content/uploads/2022/12/5.-FSA-BPM-003_BPM-for-the-Methodology-1.pdf
- Wang, B., Hughes, V., & Foulkes, P. (2019). The effect of speaker sampling in likelihood ratio based forensic voice comparison. *International Journal of Speech, Language and the Law*, *26*(1), 97–120.
- Willis, S. M., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, A., Nordgaard, A., Berger, C. E. H., Sjerps, M. J., Lucena-Molina, J. J., Zadora, G., Aitken, C. G. G., Lunt, L., Champod, C., Biedermann, A., Hicks, T. N., & Taroni, F. (2015). *ENFSI guideline for evaluative reporting in forensic science*.
- Zhang, C., & Enzinger, E. (2013). Fusion of multiple formant-trajectory-and fundamental-frequency-based forensic-voice-comparison systems: Chinese /ei1/, /ai2/, and/iau1/. In *Proceedings of Meetings on Acoustics*. Acoustical Society of America.
- Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison–female voices. *Speech Communication*, *55*(6), 796-813.

Appendix A. Place of origin of the 75 speakers in the present study

	N
ACT	2
Brisbane	3
Canberra	2
Coastal NSW	1
Country NSW	6
Country VIC	1
Ireland	1
Kempsey	1
Mackay	1
Melbourne/Sydney	1
Northern NSW	1
NSW Central Coast	1
Sydney	51
Tamworth	1
Western NSW	2
TOTAL	**75**

Appendix B. Mean frequency values (Hz) and standard deviations of F0, F1, F2 and F3 across 75 speakers in different recordings (CNV1, CNV2, INT1, INT2)




Conversation 1
F0
F1
F2
F3
No. of Speakers
SD Frequency (Hz)
Conversation 2
Interview 1
Interview 2