Evidential value of voice quality acoustics in forensic voice comparison

Ricky K. W. Chan

Speech, Language and Cognition Laboratory, School of English, University of Hong Kong

rickykwc@hku.hk

Acknowledgements

This research has been generously supported by the Hong Kong Research Grant Council Early Career Scheme (HKU Project Code: 21606918). We are grateful to Prof. Philip Rose for his invaluable feedback at the early stage of this research project, and Dr. Bruce Wang for his research assistance.

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

Abstract

Voice recordings in forensic voice comparison casework typically involve speech style mismatch and are separated by days or weeks, but studies that aim to empirically validate the evidential value of speech features rarely include systematic comparisons on contemporaneous vs. non-contemporaneous recordings and match vs. mismatch in speech style. This study addresses this gap and focuses on the acoustics of laryngeal voice quality, since voice quality has been reported to be one of the most popular and useful features for forensic voice comparison. 75 male speakers aged 18–45 were selected from a forensically-oriented database of Australian English speakers in Sydney/New South Wales. The evidential strength of a number of spectral tilt and additive noise parameters were tested under the Bayesian likelihood-ratio framework. Results show that system performance using these parameters as input were stable across 50 replications. When speech style is controlled for, VQ parameters yielded promising results and better system validity was achieved when using more VQ parameters. However, they offered limited speaker-discriminatory value when speech style mismatch is involved, and non-contemporaneous recordings only led to a small decline in performance. Overall, forensic practitioners should be cautious when using spectral tilt measures and additive noise measures as speaker discriminants in forensic casework.

Keywords: Forensic voice comparison, voice quality, speech style mismatch, non-contemporaneous recordings, likelihood-ratio

Highlights

- The evidential strength of laryngeal voice quality acoustics was assessed under the Bayesian likelihood-ratio framework.
- Speech samples from 75 male Australian English speakers in a forensically-oriented database were used.
- Spectral tilt and additive noise acoustic parameters yielded promising results when speech style is controlled for.
- Speech style mismatch and the use of non-contemporaneous recordings led to worse system performance.

1. Introduction

The task of forensic voice comparison (FVC) mostly involves comparison of voices on disputed and known samples. The disputed sample typically contains an unknown voice of an offender (possible scenarios include a masked individual recorded on CCTV footage, hoax call, ransom demand, threatening message, conversation with accomplices recorded with a covert device, intercepted phone call, etc.), and the known sample typically involves the voice of a suspect captured during a police interview (Nolan, 1983; Morrison & Enzinger, 2019). The goal of FVC is to assist the investigating authorities (e.g., police) or trier-of-fact (e.g., jury or judge) in deciding whether the known and unknown voices are from the same speaker or different speakers. With advances in speech and recording technology, recorded voices are increasingly presented as evidence in court cases (French & Stevens, 2013). More FVC research is needed in response to the anticipated mounting demand for voice comparison in forensic contexts.

One of the main goals in FVC research is to identify speech features that are useful for identifying/distinguishing voices. Empirical tests are needed to determine the evidential strengths of various speech features under conditions that are typically found in forensic casework (Morrison et al., 2021; see also Morrison & Enzinger, 2019 for a discussion on the jurisdictions and organizations that require/recommend demonstration of scientific validity and reliability of forensic evidence). Given the constraints of time and resources, it is not easy for forensic practitioners to collect new data and conduct a new validation on a case-specific basis. Therefore, anticipatory validation results can be valuable in helping forensic practitioners make informed decisions when the validation data are sufficiently similar to the conditions of the case at hand. The past two decades have witnessed a surge in research reports that involved

anticipatory validation of FVC systems[1] (i.e. validation under conditions that are expected to occur in future forensic cases; see Morrison et al. (2021) for a discussion on the factors and recording conditions that should be considered). This paper contributes to this endeavour by assessing the evidential strength of laryngeal voice quality acoustics under the likelihood-ratio (LR) framework, and how speech style mismatch and the use of non-contemporaneous recordings, which are commonly in forensic cases (Morrison et al., 2021), may affect their performance.

'Voice quality' generally refers to the quasi-permanent characteristics of one's speech running through all the sounds from the speaker (Laver, 1968). Such long-term characteristic colouring of a voice is the result of a range of *laryngeal* and *supralaryngeal* settings and activities (Biemans, 2000; Laver, 1980). The term has often been used in a narrow sense to refer to phonation types or vocal fold articulation (aka laryngeal voice quality; e.g. breathy, modal, creaky; Gordon & Ladefoged, 2001; Podesva, 2007). According to a survey on international practices among FVC experts, voice quality is considered to be one of the most useful and popular features for discriminating speakers in FVC casework, especially among forensic experts who adopted the auditory phonetic analysis approach (Gold and French, 2011). French and Stevens (2013) even argue that voice quality should always be thoroughly analyzed in FVC casework. Voice quality is typically described and analysed subjectively by experts based on auditory impressions and reported in categorical labels (e.g. tense larynx, fronted lingual body orientation), often with the Laver Vocal Profile Analysis (VPA) scheme or a modified version of such a scheme (French & Stevens, 2013; Gold & French, 2011). However, reliable auditory analysis of voice quality can be time-consuming (Nolan, 2005) and hinges on sufficient 'auditory and productive training in order to internalise the analytic perceptual

[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.

categories and associate them with their articulatory correlates' (Nolan, 2007, p.119). Also, speech signals in the incriminating/disputed speech sample are often degraded due to telephone transmission or poor recording environment. The acoustic manifestations of voice quality, especially laryngeal voice quality, may be distorted and this will adversely affect the auditory judgments of voice quality (Nolan, 2005; 2007). Furthermore, speech style variation within the same speaker will lead to different voice quality profiles. Nolan (2005, p. 402) argues that auditory judgment of voice quality with speech style mismatch or channel distortion will have to involve mental reconstruction which is 'a leap of faith on the part of the analyst'. Empirical studies are needed to demonstrate that componential auditory analysis of voice quality can be carried out in a way which compensates for the effects of channel and/or speech style mismatch (Nolan, 2007).

On the other hand, the analysis of voice quality acoustics is more transparent than impressionistic voice quality analysis, and may be incorporated in both (auditory-)acoustic-phonetic approaches or human-supervised automatic approaches to FVC, which were found to be the most popular approaches adopted by forensic experts (Gold & French, 2011; Morrison et al., 2016). Speaker-specific supralaryngeal settings (e.g. habitual palatalisation) may be assessed based on long-term formant distributions (LTFDs) (Nolan, 2007). This paper focuses on laryngeal voice quality (VQ hereafter) which has received much less research attention.

Ladefoged (1971) proposed a continuum of VQ/phonation types defined in term of the degree of opening/constriction between the arytenoid cartilages, ranging from voiceless (glottis being furthest apart), through breathy voice, to modal (regular) voicing, creaky voice and finally glottal closure. Such a classification scheme, albeit being simplistic in terms of the range of possible physiological dimensions of voice quality variation, has remained a popular model for the analysis of phonation contrasts in the world's languages (see Baken & Orlikoff, 2000; Garellek, 2019 for discussion). In terms of acoustics, glottal spreading correlates with increased

spectral tilt, whereas glottal constriction with decreased spectral tilt (Klatt & Klatt, 1990; Gordon & Ladefoged, 2001; Hanson et al., 2001; Kreiman et al., 2012). Spectral tilt is typically characterized by parameters representing harmonic amplitude differences in various harmonic-based bandwidths. For instance, H1-H2 refers to the amplitude difference between the first and second harmonic, and H1-A1 the amplitude difference between the first harmonic and the harmonic closest to the first formant. Another important dimension of describing VQ is additive noise, with typical breathy voice and creaky voice being noisier than modal voice due to the presence of aspiration noise and irregular pitch respectively (Garellek, 2019). Additive noise can be captured by the cepstral peak prominence (CPP) and harmonics-to-noise ratio (HNR) at different frequency ranges, the latter of which involves the amplitude difference between the harmonic and inharmonic components of the source spectrum (de Krom, 1993). Spectral tilt and additive noise are both important characteristics of vocal fold articulation and will be the focus of this paper; details of the acoustic parameters examined in this study are provided in the section 2.2.

Despite the purported usefulness of VQ parameters for FVC (e.g. French & Stevens, 2013; Gold & French, 2011), it is not until the past decade that empirical studies on the speaker-discriminatory power of VQ parameters, especially those using the LR framework, have been reported. Existing findings on the evidential value of VQ-based parameters appear to be mixed. Enzinger et al. (2012) assessed the performance of voice-source features extracted by GLOTTEX software package based on data from 60 female standard Mandarin speakers. They found that VQ features performed worse than automatic speaker recognition systems based on Mel-frequency cepstral coefficients (MFCCs), and that the addition of VQ features did not improve or worsen the performance of these systems. These suggest little benefit of including VQ features in FVC. But Hughes et al. (2019) noted that the poor results reported in Enzinger et al. (2012) may be due to the fact that only the nasal /n/ were examined and that some speakers

involved a small number of tokens. Vandyke et al. (2012), on the other hand, demonstrated substantial decrease in log LR cost C_{llr} when voice-source information was fused with a baseline MFCC-based system, although only 26 females were involved in the experiment. Hughes et al. (2019) explored the evidential strength of long-term harmonic and inharmonic components of laryngeal voice quality acoustics using the LR approach. Spontaneous (telephone conversation and mock police interview), contemporaneous speech data of 97 males from the DyViS database (Nolan et al., 2009) were used. The main acoustic parameters analysed were additive noise measures (cepstral peak prominence, harmonics-to-noise ratios at different frequency ranges) and spectral tilt measures (H1-A1, H1-A2, H1-A3, H1-H2, and H2-H4). All parameters were tested in four channel conditions commonly found in forensic casework: studio quality, landline telephone, and mobile phone with high or low bit rate. Gaussian Mixture Model Universal Background Model (GMM-UBM) was used to compute the scores of the input features and calibrated using logistic regression. Testing was repeated 20 times with different speakers in the development, test, and reference samples. They found that, with all voice quality measures as input, the best system performance involves a C_{llr} of 0.26 and an EER of 5.8% in high quality studio samples. Spectral tilt measures generally outperformed additive noise measures. Importantly, in general, telephone or mobile phone transmission only led to a small decline in performance, revealing that voice quality acoustics are robust to channel variation. Furthermore, the addition of voice quality information was found to improve MFCCs-based system performance, especially when transmission quality degraded.

Still, so far no study has explored the evidential strength of spectral tilt and additive noise measures using non-contemporaneous recordings. This is not surprising, as forensically-oriented databases (e.g. speech corpora that involve forensically relevant speech styles) that are made publicly available are rare, let alone ones that contain non-contemporaneous

recordings. However, FVC casework typically involves recordings that are separated by days, weeks or even months, and within-speaker variation is expected to be greater for non-contemporaneous data. To evaluate the reliability of speech features properly, at least two non-contemporaneous recordings should be used for modelling within-speaker variability in same-speaker comparisons (Enzinger & Morrison, 2012). The use of contemporaneous recordings, in contrast, may result in "underestimation of the degree of within-speaker variability" and, in turn, in "overly optimistic estimates of the degree of validity and reliability of the system" (Morrison et al., 2012, p. 157). The present study is the first to assess the evidential strength of laryngeal voice quality acoustics using non-contemporaneous recordings. Another goal of the present study is to determine the impact of speech style mismatch, which is commonly found in forensic casework, on the speaker-discriminatory power of VQ acoustic parameters. This was achieved by including both same-style comparison and different-style comparison and determines how mismatch in speech style may affect the evidential strength of VQ parameters.

In sum, the present study aims to determine the evidential strength of spectral tilt and additive noise measures under the LR framework and compare system performance using 1) contemporaneous vs. non-contemporaneous recordings; and 2) match vs. mismatch in speech style. It is expected that both speech style mismatch and non-contemporaneous recordings will lead to greater within-speaker variation and hence worse system performance. Still, assessing the relative magnitude of these factors will provide valuable information for forensic practitioners when incorporating VQ parameters in future forensic casework.

2. Methods

2.1 Corpus

The speech data came from a forensically-oriented database of 552 Australian English speakers (332 females and 231 males at the time of writing) (Morrison et al., 2015)[2]. Each speaker was recorded on one to three or more occasions based on the protocol proposed in Morrison et al. (2012). In each recording session, each speaker completed three speaking tasks: casual telephone conversation with a friend/colleague (CNV), fax information exchange over the telephone, and pseudo-police interview (INT). For speakers recorded on more than one occasion, the time interval between the recording sessions was about two weeks. The recordings were saved in a high-quality format with noise and cross-talk manually removed.

75 speakers were chosen from the database based on the following considerations: age, gender, regional background, and the availability of non-contemporaneous recordings. Non-contemporaneous recordings were defined here as recordings separated by at least a one-week interval. We picked male speakers only as most forensic casework involves males and males are more commonly involved in crime than females (Steffensmeier & Allan, 1996). Besides, only speakers who were recorded on more than one occasion were chosen given the importance of using non-contemporaneous recordings in the evaluation of FVC systems (Morrison et al., 2012). Then we strived to control for speakers' age and regional background as far as possible. To this end, 75 male speakers aged between 18 and 45 were selected and most of them were from Sydney and other areas within the state of New South Wales. The breakdown of the speakers' regional background can be found in Appendix A. For each speaker, four recordings—the CNV and INT tasks recorded in two separate sessions (i.e. CNV1, CNV2, INT1, INT2)—were analyzed. We focused on these two tasks because the speaking styles involved are typically found in forensic casework (Morrison et al., 2012).

[2] Details of the database can be found at http://databases.forensic-voice-comparison.net/#australian_english_500.

2.2 Feature extraction and parameterization

Vowel-only portions of the recordings were manually segmented and labelled in Praat (Boersma, Weenink, 2014). Approximately 33 seconds of net vocalic material was extracted per speaker per recording. VQ parameters reported in Hughes et al. (2019) were selected so that the findings will be comparable. They were extracted using VoiceSauce (Shue, 2011) with a 20ms window length and 10ms window shift.

*H1-H2 and H2-H4***.** These are the amplitude differences between the first and second and second and fourth harmonics respectively.

H1-A1, H1-A2, and H1-A3: these are the amplitude differences between the first harmonic and the spectral magnitude at the first, second and third formant respectively.

The harmonic/spectral amplitudes were corrected for formant frequencies and bandwidths (see Shue et al., 2010 for details). In general, the higher values of these measure, the greater the spectral tilt which suggests a higher degree of glottal spreading due to breathiness. Vice versa for glottal constriction due to creakiness. These five measures were also combined as the 'spectral tilt' which was analyzed separately.

Cepstral peak prominence (CPP). a measure of cepstral peak amplitude normalized for overall amplitude (Hillenbrand et al., 1994). In principle, modal phonation has well-defined periodic waves that result in larger cepstral peaks, while breathy phonation is likely to have less well-defined ones and lower cepstral peaks. A larger CPP value indicates a more modal voice, whereas a smaller CPP value a breathier voice.

Harmonic-to-noise ratio (HNR). HNR captures the spectral noise level and is positively correlated to the degree of perceived breathiness (de Krom, 1993). The HNR was extracted over 0-500Hz (HNR05), 0-1500Hz (HNR15), 0-2500Hz (HNR25), and 0-3500Hz (HNR35) respectively, resulting in four separate measures.

These five measures were also combined as the 'additive noise' which was analyzed separately.

Outliers were removed as they were deemed unrepresentative of the speakers' typical long-term voice quality characteristics. They were defined as data points that are three median absolute deviations away from the overall median, as opposed to the commonly used "the mean plus or minus three standard deviations" approach because the mean and standard deviation are strongly influenced by outliers (see Leys et al., 2013 for discussion). These voice quality features then served as the input for score generation and subsequent LR computation.

2.3 Statistical analysis

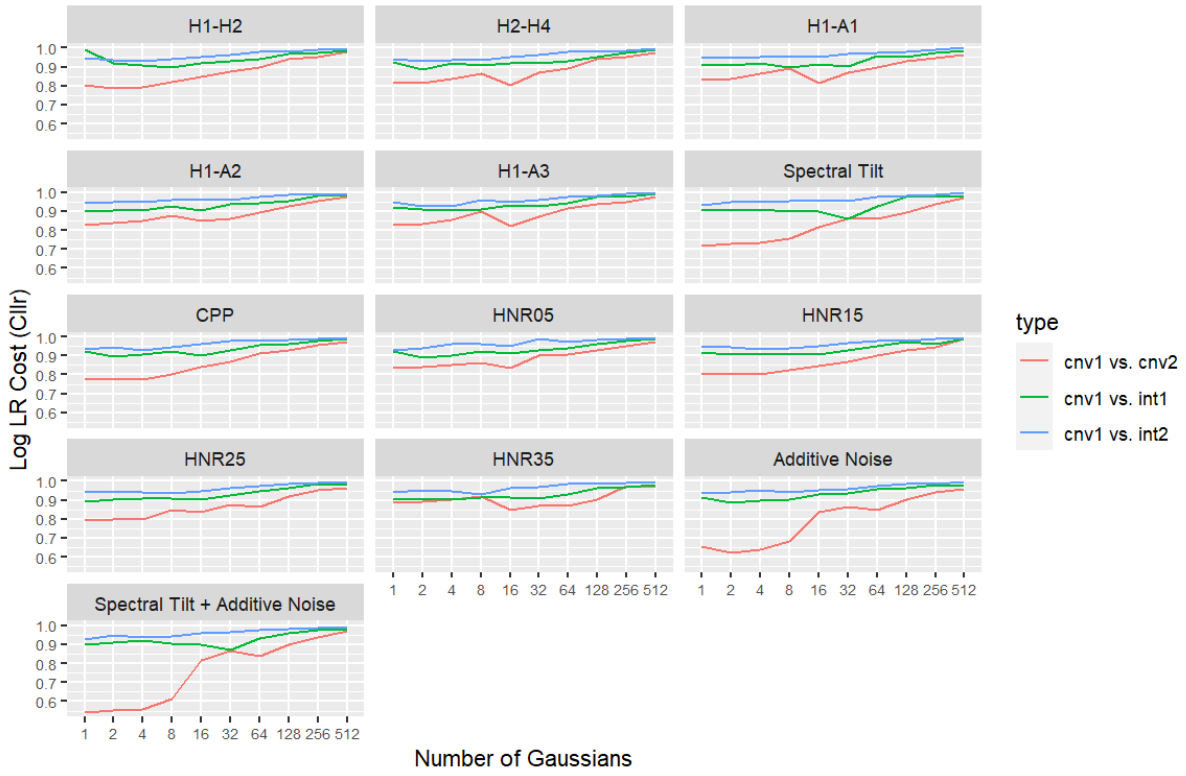
To assess the evidential strength of voice quality parameters, the Gaussian Mixture Model-Universal Background Model (GMM-UBM) (Reynolds et al., 2001) was used to generate same-speaker and different-speaker comparison scores in MATLAB. GMMs were fitted with varying number of Gaussians (1, 2, 4, 8, 16, 32, 64, 128, 256, 512) in order to test which option(s) would fit the data better (Jessen, 2021). Other statistical analysis was conducted in R (R Development Core Team, 2008).

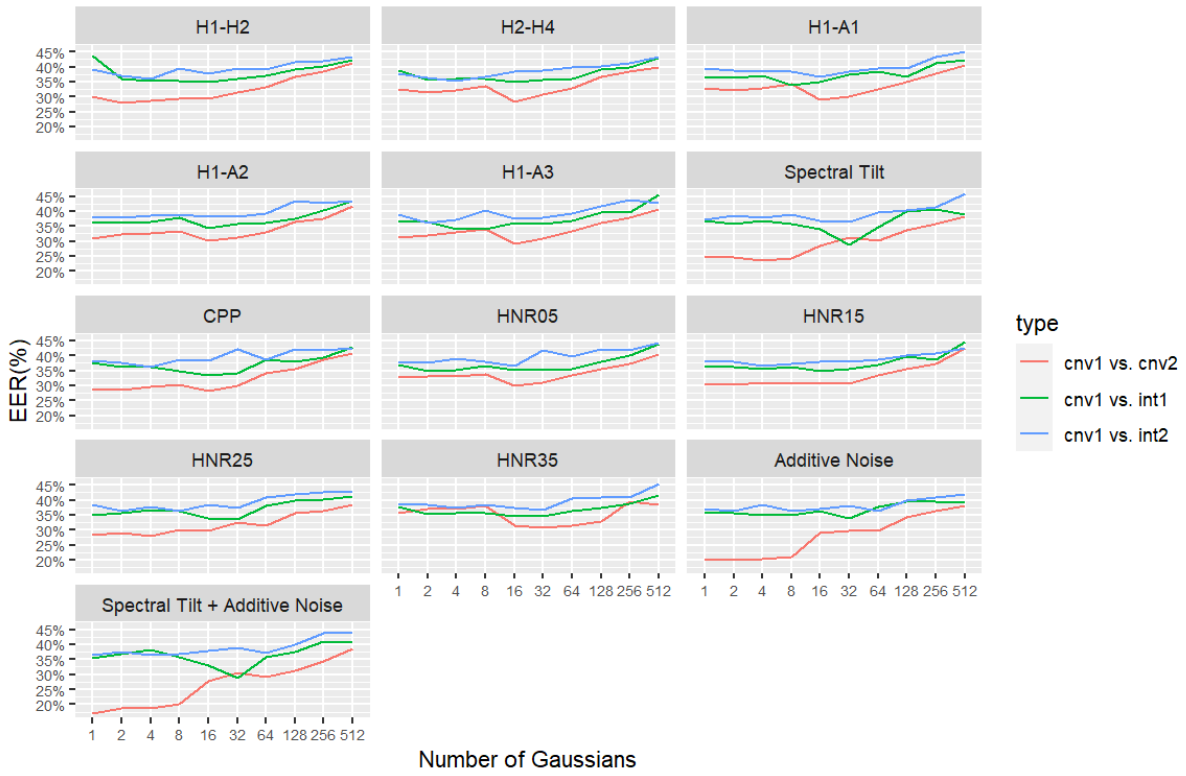
Calibration was conducted using logistic regression (Brummer et al., 2007). The 75 speakers were randomly assigned to one of the three datasets: training, test, or reference set (25 speakers in each set). The procedure above was replicated 50 times with random allocations of speakers in the training, test, and reference sets, as it has been demonstrated that the reliability of system performance hinges on the speaker samples involved (Wang et al., 2019). To test the effects of speech style mismatch and the use of non-contemporaneous recordings, the evidential strength of VQ features were tested using three different sets of speech data: 1) CNV1 vs.

CNV2 (same speech style, non-contemporaneous recordings); 2) CNV1 vs. INT1 (different speech styles, contemporaneous recordings); and 3) CNV1 vs. INT2 (different speech styles, non-contemporaneous recordings). System validity was evaluated based on two common metrics that are widely used in the FVC literature: equal error rate (EER) and log-LR cost (C_{llr}) (Brümmer & du Preez, 2006). The lower the C_{llr}/EER value, the better the system performance. C_{llr} values close to or greater than 1 imply limited evidential value of the input parameter(s).

3. Results

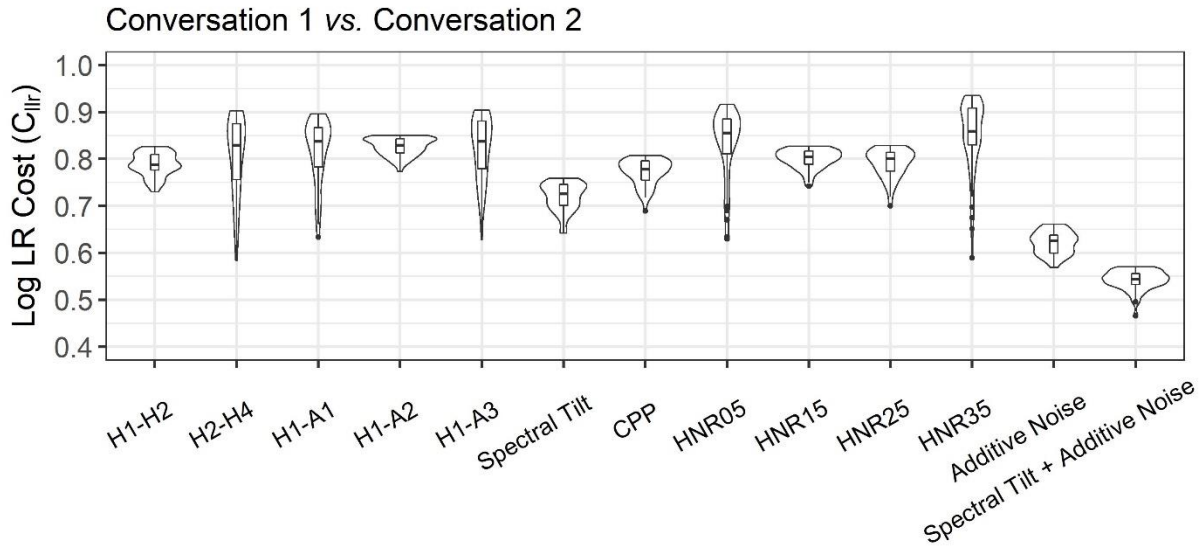
Mean values and standard deviations of individual VQ parameters across the 75 speakers in the four speech samples (CNV1, CNV2, INT1, INT2) are included in Appendices B and C. Figures 1 and 2 show the mean C_{llr} and EER values with varying number of Gaussians. Despite some fluctuations, it appears that in general, increasing the number of Gaussians from 1 to 512 led to worse system performance: the mean C_{llr} and EER values became gradually higher. This is in line with the results based on long-term formant analysis conducted by Jessen (2021). Best performance is often achieved with just one Gaussian. This suggests that in general a small number of Gaussians is sufficient for modelling the score distribution of the voice quality parameters under investigation.

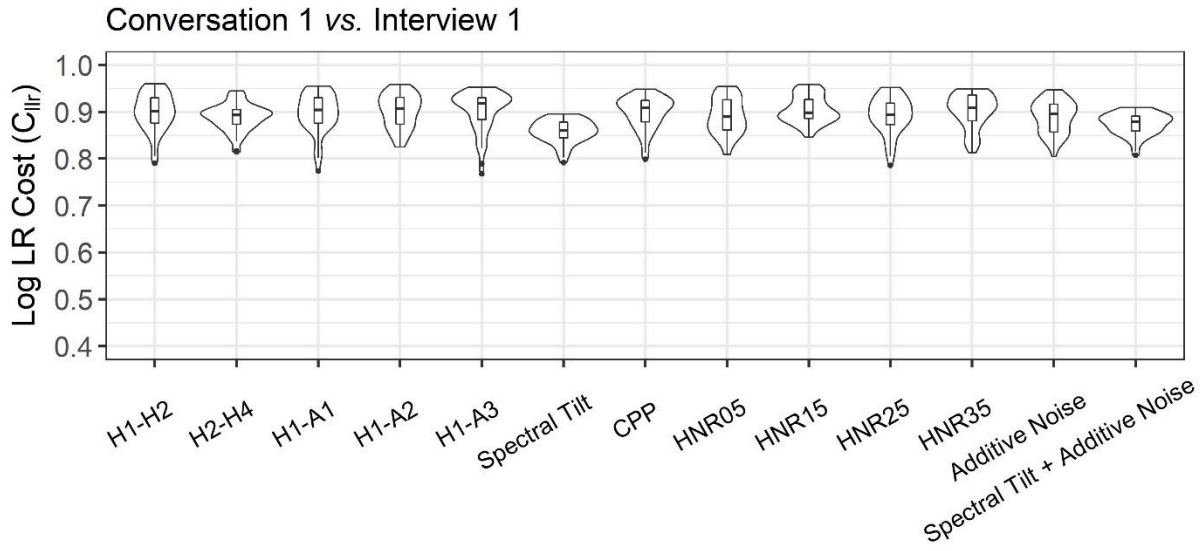


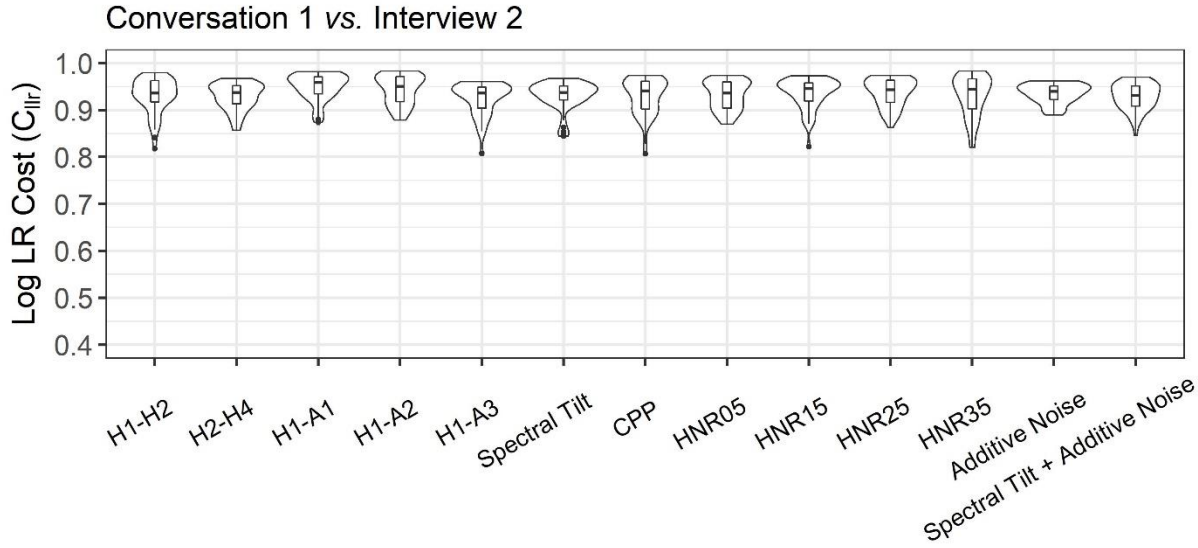


Figures 1 and 2: Mean Cllr and EER values of systems based on individual VQ parameters or combinations thereof, with increasing number of Gaussians for CNV1 vs. CNV2, CNV1 vs. INT1, and CNV1 vs. INT2 respectively.

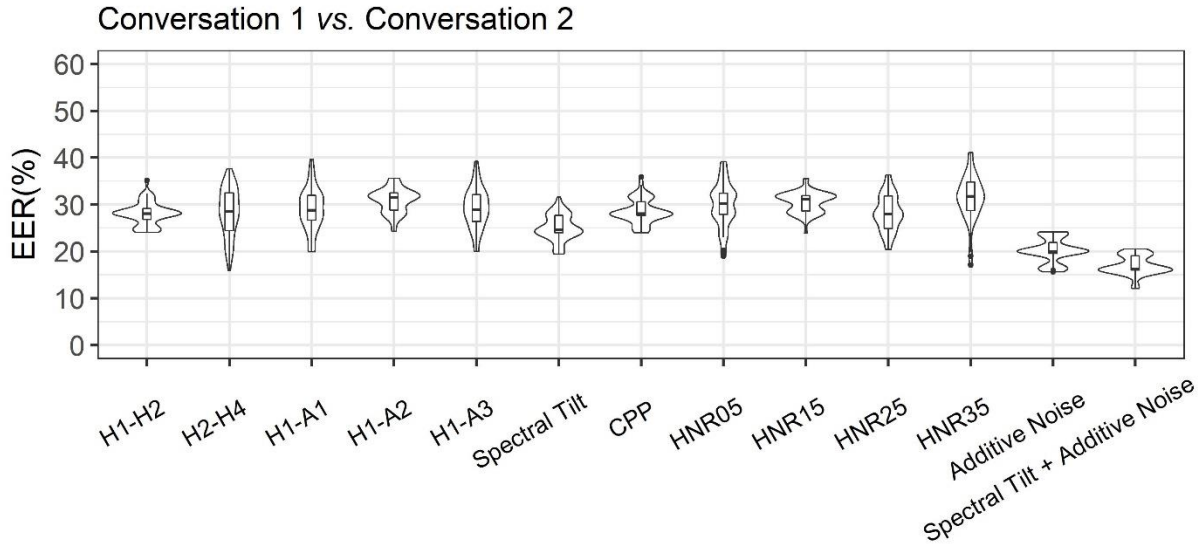
The subsequent results are based on the number of Gaussians that generated the best results for the individual VQ parameters and combinations thereof. Figures 3 to 5 show the distributions of C_{llr} values for individual VQ parameters (i.e. single-parameter systems), combined spectral tilt measures (5 parameters) and combined additive noise measures (5 parameters) for CNV1 vs. CNV2, CNV1 vs. INT1, and CNV1 vs. INT2. Figures 6 to 8 show their EER values. Tables 1 to 3 shows the minimum, maximum, mean, and standard deviation of their C_{llr} and EER values.

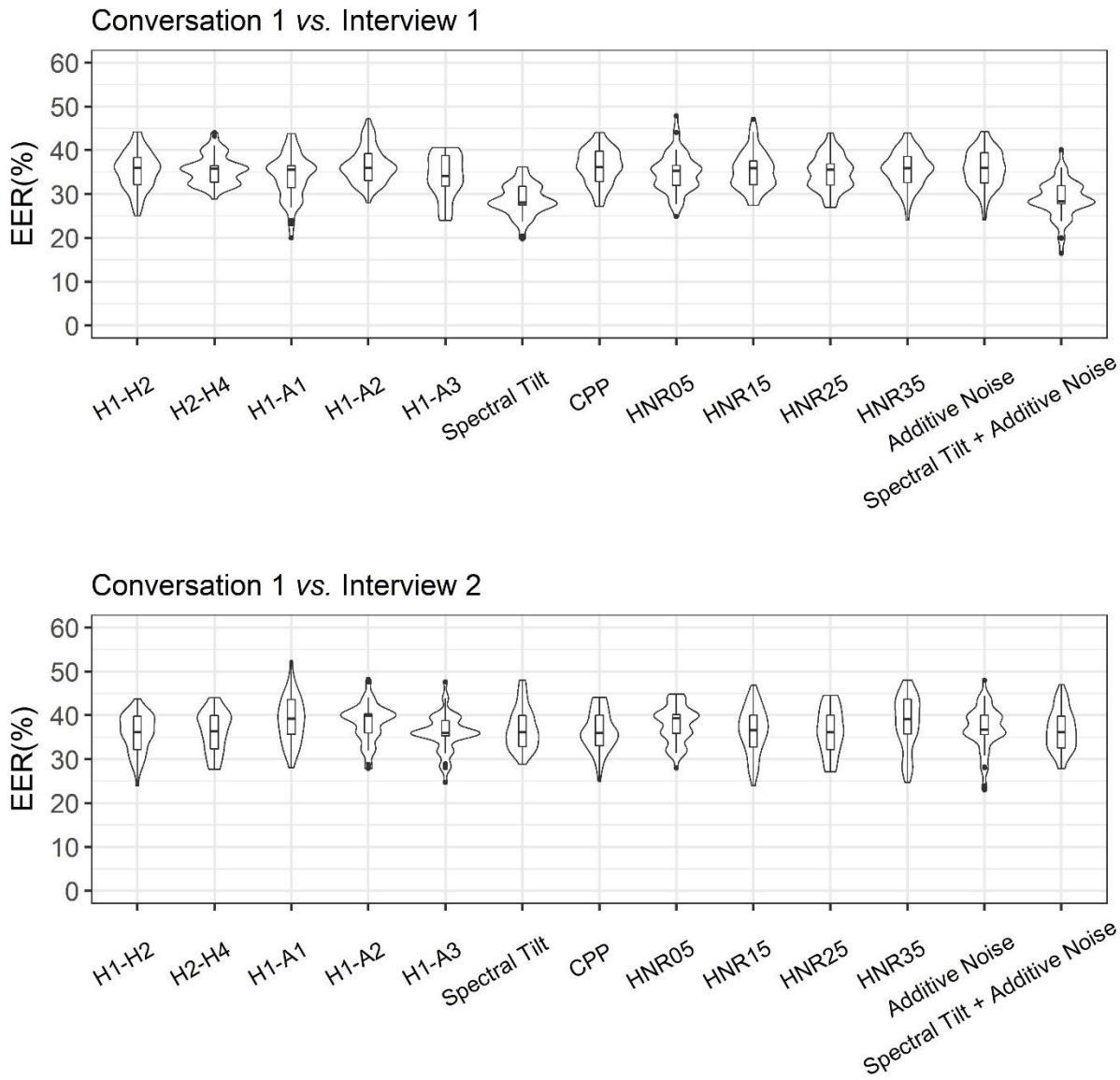






Figures 3 to 5: Distribution of C_{llr} values using individual VQ parameters and combinations thereof in CNV1 vs. CNV2, CNV1 vs. INT1, and CNV1 vs. INT2 respectively.





Figures 6 to 8: Distribution of EER (%) values using individual VQ parameters and combinations thereof in CNV1 vs. CNV2, CNV1 vs. INT1, and CNV1 vs. INT2 respectively.

CNV1 vs. CNV2

	C_{llr}				EER (%)			
VQ parameter	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
H1 - H2 (2)	0.73	0.83	0.79	0.02	24.08	35.17	28.06	2.42
H2 - H4 (16)	0.58	0.90	0.81	0.08	15.92	37.67	28.46	5.18
H1 - A1 (16)	0.63	0.90	0.82	0.07	19.92	39.67	29.01	4.61
H1 - A2 (1)	0.77	0.85	0.83	0.02	24.25	35.58	30.90	2.59
H1 - A3 (16)	0.63	0.90	0.82	0.07	20.00	39.25	29.27	4.34
Spectral tilt (1)	0.64	0.76	0.72	0.03	19.42	31.58	25.05	2.80
CPP (2)	0.69	0.81	0.77	0.03	24.00	35.92	28.57	2.61
HNR05 (16)	0.63	0.92	0.83	0.07	19.00	39.17	29.90	4.83
HNR15 (2)	0.74	0.83	0.80	0.02	23.92	35.50	30.49	2.22

HNR25 (1)	0.70	0.83	0.79	0.03	20.42	36.33	28.22	3.90
HNR35 (16)	0.59	0.94	0.85	0.08	17.08	41.08	31.41	4.86
Additive noise (2)	0.57	0.66	0.62	0.02	15.67	24.17	20.16	2.53
Spectral tilt + additive noise (1)	0.47	0.57	0.54	0.02	12.08	20.50	16.98	2.02

CNV1 vs. INT1

	C_{llr}				EER(%)			
VQ parameter	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
H1 - H2 (8)	0.79	0.96	0.90	0.04	25.00	44.17	35.37	4.34
H2 - H4 (2)	0.81	0.94	0.89	0.03	28.83	44.00	35.57	3.34
H1 - A1 (8)	0.77	0.95	0.90	0.04	20.00	43.83	33.98	5.16
H1 - A2 (1)	0.83	0.96	0.90	0.04	28.00	47.25	36.53	4.00
H1 - A3 (4)	0.77	0.95	0.90	0.04	24.00	40.58	33.98	4.96
Spectral tilt (32)	0.79	0.90	0.86	0.02	19.75	36.17	28.70	3.56
CPP (2)	0.80	0.95	0.90	0.03	27.17	44.00	36.33	4.05
HNR05 (2)	0.81	0.95	0.89	0.04	24.83	47.83	34.83	4.48
HNR15 (4)	0.85	0.96	0.91	0.03	27.42	47.08	35.40	4.38
HNR25 (1)	0.79	0.95	0.89	0.04	27.00	43.92	34.87	3.99
HNR35 (4)	0.81	0.95	0.90	0.04	24.08	43.92	35.57	4.05
Additive noise (2)	0.80	0.95	0.89	0.03	24.08	44.25	35.79	4.27
Spectral tilt + additive noise (32)	0.81	0.91	0.87	0.02	16.42	40.08	28.88	4.29

CNV1 vs. INT2

	C_{llr}				EER(%)			
VQ parameter	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
H1 - H2 (4)	0.82	0.98	0.93	0.04	23.92	43.75	35.95	4.35
H2 - H4 (2)	0.86	0.97	0.93	0.03	27.67	44.00	36.46	4.63
H1 - A1 (2)	0.87	0.98	0.95	0.03	28.08	52.50	38.87	5.31
H1 - A2 (1)	0.88	0.98	0.94	0.03	27.92	48.17	38.33	4.48
H1 - A3 (2)	0.81	0.96	0.93	0.03	24.75	47.58	36.01	4.22
Spectral tilt (1)	0.84	0.97	0.93	0.03	28.83	48.00	37.09	4.71
CPP (4)	0.81	0.97	0.93	0.04	24.92	44.08	36.40	4.47
HNR05 (1)	0.87	0.97	0.93	0.03	28.00	44.83	37.93	4.13
HNR15 (4)	0.82	0.97	0.94	0.03	24.00	46.83	36.60	5.27
HNR25 (8)	0.86	0.97	0.94	0.03	27.08	44.50	36.51	5.14
HNR35 (8)	0.82	0.98	0.93	0.04	24.75	48.00	38.35	6.11
Additive noise (1)	0.89	0.96	0.94	0.02	23.00	47.92	36.98	5.04
Spectral tilt + additive noise (1)	0.85	0.97	0.93	0.03	27.83	47.00	36.65	4.68

Tables 1 to 3: descriptive statistics of C_{llr} and EER values across 50 replications with VQ parameters as input in CNV1 vs. CNV2, CNV1 vs. INT1, and CNV1 vs. INT2 respectively. The numbers in the parentheses denote the number of Gaussians that yielded the best results.

In general, all the input parameters yielded a rather small standard deviation in C_{llr} (less than 0.1) and EER (mostly less than 5%) values across the 50 replications, suggesting that system performance using these parameters as input were generally stable (i.e. high system reliability). Specific results from the three sets of recordings are summarized below.

CNV1 vs. CNV2. When using non-contemporaneous recordings and controlled for speech style, individual additive noise parameters appeared to carry some speaker-discriminatory power, with the best C_{llr} values ranging from 0.58 (H2-H4) to 0.77 (H1-A2) and EER values from 15.92% to 24.25%, and mean C_{llr} values around 0.8. Much more promising results were achieved for the combined spectral tilt and the combined additive noise measures, with mean C_{llr} values of 0.72 and 0.62 respectively. The best replications yielded a C_{llr} of 0.64 and an EER of 19.42%, and a C_{llr} of 0.57 and an EER of 15.67% correspondingly. Using spectral tilt + additive noise parameters as input led to promising results, with a mean C_{llr} value of 0.54 and the lowest C_{llr} and EER being 0.47 and 12.08% respectively.

CNV1 vs. INT1. When using contemporaneous recordings with mismatch in speech style, in general poorer results were obtained when compared with CNV1 vs. CNV2. Individual VQ parameters had mean C_{llr} values of around 0.9 and optimal C_{llr} values ranging from 0.77 to 0.85 and EER values from 20% to 28.83%. Unlike in CNV1 vs. CNV2, the combined spectral tilt, the combined additive noise measures, and spectral tilt + additive noise measures did not seem to yield better results, with the highest mean C_{llr} value of 0.86 and the best replications producing a C_{llr} of 0.79 and an EER of 19.75%. This suggest that speech style mismatch has a more detrimental effect on speaker-discriminatory performance than non-contemporaneous recordings.

CNV1 vs. INT2. The comparison of these two datasets most closely reflects real-life forensic situation where both speech style mismatch and non-contemporaneous recordings are

involved, and even worse results were found. Individual VQ parameters appear to provide limited information for discriminating speaker. The best replications came from H1-A3 and CPP, with a C_{llr} of 0.81 and an EER of around 24%. Similarly, the combined spectral tilt, the combined additive noise measures, and spectral tilt + additive noise measures did not appear to perform better. Using all the 10 VQ parameters only led to a C_{llr} of 0.85 and an EER of 27.83% in the best replication. All VQ parameters and their combinations thereof yielded mean C_{llr} values of 0.93 to 0.95, suggesting limited speaker-discriminatory performance in CNV1 vs. INT2.

4. Discussion

One of the FVC research goals is to evaluate the evidential value of speech features, and it is crucial to test these features under forensically relevant conditions as a kind of anticipatory validation for future forensic casework (Morrison et al., 2021). The present study sought to determine the evidential strength of VQ acoustic parameters under the LR framework. The effects of speech style mismatch and the use of non-contemporaneous recordings—which are commonly found in forensic casework—were also tested, as existing FVC studies that involved systematic comparison between contemporaneous and non-contemporaneous recordings or match vs. mismatch in speech style are limited.

The present study examined the same VQ parameters as in Hughes et al. (2019), but the findings differ in a number of ways. Hughes et al. tested these parameters only with contemporaneous speech data that involved speech style mismatch (conversation vs. interview), which was similar to the CNV1 vs. INT1 comparison in the present study. They found that the combined spectral tilt measure and the combined additive noise measure carry considerable speaker-discriminatory information, and system validity based on these input features was only

slightly affected by channel mismatch and deterioration of recording quality. However, our findings are less promising and suggest that these measures are of limited evidential value despite the similarities in the speech data involved (tasks: conversation vs. police interview; studio recording quality). This might be attributable to the methodological differences between the two studies. First, Hughes et al. analyzed more speakers than the present study (97 vs. 75). However, it has been shown that stable LR output can be achieved with 20 speakers in the training set (e.g. Hughes, 2017). The present study involved 25 training speakers, 25 test speakers and 25 reference speaker, and system performance is generally stable as reflected by low standard deviation in C_{llr} values across 50 replications. Second, Hughes et al. involved longer vocalic material per recording (approximately 60s vs. 33s). However, a recent study on long-term formant analysis by Jessen (2021) revealed that reasonable system performance could be achieved with merely 10s pure vowels. Whilst the amount of speech data needed for reliable long-term voice quality acoustic analysis in FVC remains an empirical question, 33s of vocalic material should be sufficient for capturing speakers' long-term voice quality characteristics in the present study. Third, the two studies involved different varieties of English (southern British English (SBE) vs. Australian English). The performance differences suggest that the evidential strength of VQ parameters may be language/variety-specific, and cautions should be exercised when generalising the results to other languages. Nonetheless, neither breathiness nor creakiness is contrastive in Australian English or SBE, and there appears to be no socially-conditioned phonation variation in the conversation and mock police interview tasks among the male speakers in our dataset. The linguistic difference between the two varieties of English is not expected to play a major role here. The observed performance discrepancies may be due to the nature of the datasets. For example, it might be the case that within-speaker differences stemming from speech style variation were greater for our data, or that Australian English speaker in our data are more similar-sounding than the SBE speakers

analyzed in Hughes et al. (2019). Systematic comparison between the two datasets have to be conducted in order to test these speculations.

In addition, the present study involves two more sets of comparisons and results show clear effects of speech style mismatch and non-contemporaneous recordings on the performance of VQ acoustic parameters. In CNV1 vs. CNV2 (same speech style, non-contemporaneous recordings), individual VQ parameters performed reasonably well in distinguishing speakers, and no single VQ parameter seems to outshine the others. System performance improved considerably when the five spectral tilt measures or the five additive noise measures were combined as input, and there is no clear performance difference between these two dimensions of voice quality acoustics. This shows that spectral tilt and additive noise measures have similar speaker-discriminatory value. Even better performance was achieved for spectral tilt + additive noise, suggesting that spectral tilt measures and additive noise measures offer some degree of complementary information for distinguishing speakers. Overall, these results are in line with previous findings that maximal speaker discrimination is often achieved by using a combination of complementary speech features rather than single features, and that systems with more non-overlapping parameters as input may lead to higher system validity (e.g. Chan, 2020; Hughes et al., 2016; 2019).

However, results are much less promising when speech style mismatch is involved (CNV1 vs. INT1 or INT2). Individual VQ parameters performed rather poorly in distinguishing speakers as reflected by the high C_{llr} and EER values, and no single VQ parameter appears to consistently stand out from the others. Unlike in CNV1 vs. CNV2, system validity did not show clear improvement even when more parameters are added to the system (spectral tilt, additive noise, spectral tilt + additive noise). This reveals speech style mismatch has clear negative effects on the evidential strength of VQ acoustic parameters and might have overridden the potential benefits of using more parameters. The comparison of results between CNV1 vs.

INT1 and CNV1 vs. INT2 reveals that non-contemporaneous recordings only led to slight worse system performance. It can be concluded that the speaker-discriminatory performance of VQ parameters is generally affected more by speech style mismatch than non-contemporaneous recording. The detrimental effects of speech style mismatch found in the present study highlight the fact that aspects of voice quality effects are not only ingrained by habits and bound by physiological limits, but may also be voluntarily manipulated in different speaking styles (Nolan, 2007). The within-speaker variation involved may render two recordings of the same speaker sound very different. On the other hand, the negative effects of non-contemporaneous recordings suggest that a speaker voice quality may change from occasion to occasion which can be random or conditioned (Morrison et al., 2012), but such effects appear to be small. Still, this study only involved informal conversation and mock police interview and non-contemporaneous recordings with a time interval of approximately two weeks. It should be noted that in forensic cases variations in speech style and time interval between recordings are much more variable. Future research should investigate the performance of VQ parameters in other speech styles typically found in forensic casework (e.g. speech under various emotional states), and with speech samples separated by different time periods.

More importantly, with both speech style mismatch and non-contemporaneous recordings (i.e. CNV1 vs. INT2) as in typical forensic casework, VQ acoustic parameters performed rather poorly and appear to bear little speaker-discriminatory value in actual forensic cases, as least when Australian English is concerned. At a broader level, this seems to be at odds with the claim among some forensic experts that voice quality is one of the most useful features for FVC casework (e.g. French & Steven, 2013; Gold & French, 2011). Nonetheless, the present study only analyzed laryngeal voice quality, with a focus on a number of spectral tilt and additive noise parameters. These parameters correspond to auditory VQ labels such as

creaky voice and breathy voice, but they only constitute a portion of the voice quality characteristics normally analysed in actual forensic casework. French and Steven (2013) note that in typical voice quality analysis using a version of the Laver VPA scheme, around 38 speech features and vocal tract settings (e.g. creaky voice, nasalisation, fronted lingual body orientation) may be analyzed auditorily, and each of the features is assigned a score on a scalar degree. A comprehensive analysis of the acoustic correlates of all these features/settings is necessary in order to fully evaluate the value of voice quality analysis in forensic casework. Still, whilst impressionistic vocal profile analysis analysis may sometimes be corroborated with spectrographic evidence, the human perception and subjective judgments involved are non-transparent, non-reproducible and may be susceptible to cognitive bias and human errors (Edmond et al., 2007; Morrison, 2022). Also, although auditory VPA is considered very useful in a FVC task, it requires formal training (French & Steven, 2013) and performance variability across analysts should not be overlooked. On the other hand, the analysis of VQ acoustics in this study involves quantitative measurements, statistical models, and calibration using data that are similar to forensic conditions. This is in line with the ongoing paradigm shift in the evaluation of forensic evidence across different fields of forensic science—moving from methods that hinge on subjective judgment and human perception, to quantitative analysis with statistical modelling of data relevant to forensic scenarios (Saks & Koehler, 2005; Morrison, 2009; 2022). Emphasis is also placed on transparency, replicability, and the reporting of accuracy and error rates. It is hoped that the findings will help transform the categorical auditory-based labels in VPA into continuous acoustic variables whose evidential strength can be easily assessed using the LR framework for forensic casework. Future research should assess other acoustic parameters of voice quality under forensically-relevant conditions, and how voice quality parameters may be combined with other features for capturing maximal speaker-discriminatory information.

5. Conclusion

Given the importance of empirically validating speech features to be used in FVC casework, the present study investigated the evidential strength of a range of spectral tilt and additive noise parameters (and their combinations) under the likelihood-ratio framework. Contrary to the finding by Hughes et al. (2019) and the claim that voice quality is one of the most useful features for FVC casework (Gold & French, 2011; French & Stevens, 2013), we found that these laryngeal voice quality parameters generally offer little speaker-discriminatory value when both speech style mismatch and non-contemporaneous recordings were involved. Forensic analysts should be cautious when using spectral tilt measures and/or additive noise measures as speaker discriminants, and it is hoped that our findings will help them make informed decisions given the circumstances of the case at hand in the future.

Reference

Baken, R. J. & Orlikoff, R. F. (2000). *Clinical Measurement of Speech and Voice*. San Diego: Singular Publishing Group.

Biemans, M. (2000). *Gender Variation in Voice Quality*. Utrecht, The Netherlands: Lot.

Boersma, P., Weenink, D., 2014. Praat: Doing phonetics with computers. Retrieved from www.praat.org.

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. *Proceedings of IEEE Transactions on Audio, Speech, and Language*, 15(7), 2072–2084.

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2-3), 230–275.

Chan, R. (2020). Speaker discrimination: Citation tones vs. coarticulated tones. *Speech Communication*, *117*, 38–50. https://doi.org/10.1016/j.specom.2019.06.006

de Krom, G. (1993). A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals. *Journal of Speech, Language, and Hearing Research*, *36*(2), 254–266. https://doi.org/10.1044/jshr.3602.254

Edmond, G., Towler, A., Growns, B., Ribeiro, G., Found, B., White, D., … Martire, K. (2017). Thinking forensics: Cognitive science for forensic practitioners. *Science & Justice*, *57*(2), 144–154. https://doi.org/10.1016/j.scijus.2016.11.005

Enzinger, E., & Morrison, G. S. (2012). The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology, Sydney.* Australia.

Enzinger, E., Zhang, C., Morrison, G.S. 2012. Voice source features for forensic voice comparison – an evaluation of the GLOTTEX software package. *Proc. Odyssey*. Singapore, 78–85.

French, P., & Stevens, L. (2013). Forensic speech science. In M. Jones & R.-A. Knight (Eds.), *The Bloomsbury Companion to Phonetics* (pp. 183-197). London: Bloomsbury.

Garellek, M. (2019). The phonetics of voice. In W. Katz & P. Assmann (Eds.), *The Routledge handbook of phonetics* (pp. 75-106). Abingdon-on-Thames, UK: Routledge.

Garellek, M., Keating, P., Esposito, C. M., & Kreiman, J. (2013). Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America, 133*(2), 1078-1089.

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech Language and the Law, 18*(2), 293-307. https://doi.org/10.1558/ijsll.v18i2.293

Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics, 29*(4), 383–406. https://doi.org/10.1006/jpho.2001.0147

Hanson, H. M., Stevens, K. N., Kuo, H.-K. J., Chen, M. Y., & Slifka, J. (2001). Towards models of phonation. *Journal of Phonetics, 29*(4), 451–480. https://doi.org/10.1006/jpho.2001.0146

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic Correlates of Breathy Vocal Quality. *Journal of Speech, Language, and Hearing Research, 37*(4), 769–778. https://doi.org/10.1044/jshr.3704.769

Hughes, V., Cardoso, A., Foulkes, P., French, J. P., Harrison, P. & Gully, A. (2019). Forensic voice comparison using long-term acoustic measures of voice quality. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS).* Melbourne, Australia.

Hughes, V., Foulkes, P. & Wood, S. (2016). Formant dynamics and durations of um improve the performance of automatic speaker recognition systems. *Proceedings of the 16th Australasian Conference on Speech Science and Technology (ASSTA)*. University of Western Sydney, Australia, 249-252.

Jessen, M. (2021). MAP Adaptation Characteristics in Forensic Long-Term Formant Analysis. *Interspeech*, 411-415.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87*(2), 820–857. https://doi.org/10.1121/1.398894

Kreiman, J., & Gerratt, B. R. (2012). Perceptual interaction of the harmonic source and noise in voice. *The Journal of the Acoustical Society of America, 131*(1), 492–500. https://doi.org/10.1121/1.3665997

Kreiman, J., Shue, Y.-L., Chen, G., Iseli, M., Gerratt, B. R., Neubauer, J., & Alwan, A. (2012). Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *The Journal of the Acoustical Society of America, 132*(4), 2625–2632.

Ladefoged, P. (1971). *Preliminaries to Linguistic Phonetics*. Chicago: University of Chicago Press.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. New York: Cambridge University Press.

Laver, J. D. M. (1968). British Journal of Disorders of Communication. *International Journal of Language & Communication Disorders*, *3*(1), 43–54. https://doi.org/10.3109/13682826809011440

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. https://doi.org/10.1016/j.jesp.2013.03.013

Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech Language and the Law*, *16*(1), 31-57.

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. (2022). Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science. *Forensic Science International: Synergy*, *5*, 100270. https://doi.org/10.1016/j.fsisyn.2022.100270

Morrison, G.S., Enzinger, E. (2019). Introduction to forensic voice comparison. In Katz, W.F., Assmann, P.F. (Eds.) *The Routledge Handbook of Phonetics* (pp. 599–634). Abingdon, UK: Taylor & Francis.

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., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, *44*(2), 155–167. https://doi.org/10.1080/00450618.2011.630412

Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Goemans Dorny, C. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, *263*, 92–100. https://doi.org/10.1016/j.forsciint.2016.03.044

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.

Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

Nolan, F. (2005). Forensic speaker identification and the phonetic. In W. J. Hardcastle & J. M. Beck (Eds), *A Figure of Speech: A Festschrift for John Laver*, 385-411.

Nolan, F. (2007). Voice quality and forensic speaker identification. *Govor, 24*(2), 111-128.

Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics*, *11*(4), 478–504. https://doi.org/10.1111/j.1467-9841.2007.00334.x

R Development Core Team (2008). R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

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

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, 309(5736), 892-895.

Shue, Y. L., Chen, G., & Alwan, A. (2010). On the interdependencies between voice quality, glottal gaps, and voice-source related acoustic measures. In *Eleventh Annual Conference of the International Speech Communication Association*.

Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. (2011). VoiceSauce: A program for voice analysis. *Proceedings of the ICPhS XVII*, 1846-1849.

Steffensmeier, D., & Allan, E. (1996). Gender and Crime: Toward a Gendered Theory of Female Offending. *Annual Review of Sociology*, *22*(1), 459–487. https://doi.org/10.1146/annurev.soc.22.1.459

Vandyke, D., Wagner, M., Goecke, R., & Chetty, G. (2012). Speaker identification using glottal-source waveforms and support-vector-machine modelling. *Proceedings of Speech Science and Technology*, 49-52.

Wang, B. X., 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. https://doi.org/10.1558/ijsll.38046

Wang, B. X., Hughes, V., & Foulkes, P. (2022). The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison. *Speech Communication, 138*, 38–49. https://doi.org/10.1016/j.specom.2022.01.009

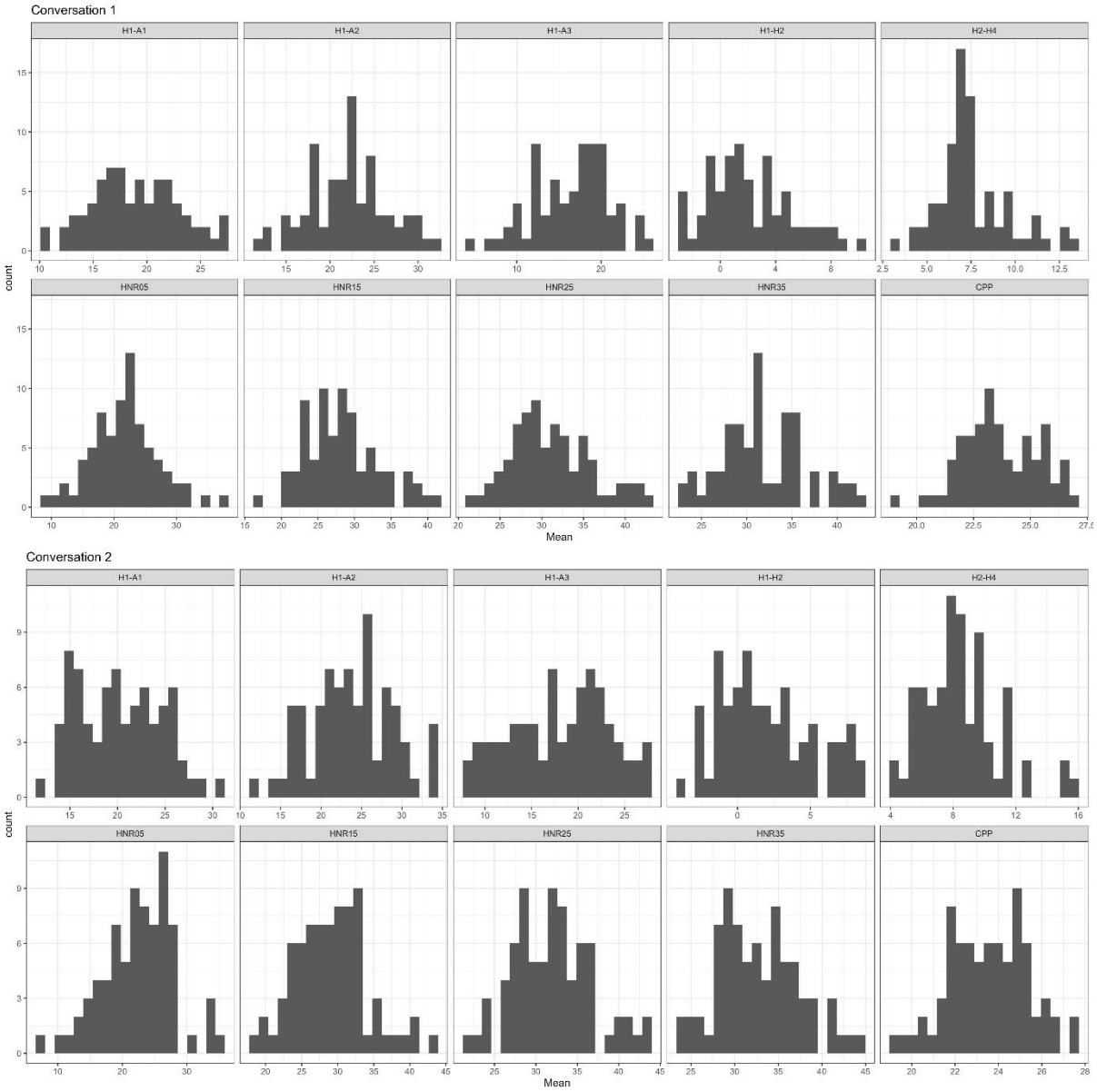
Appendix A

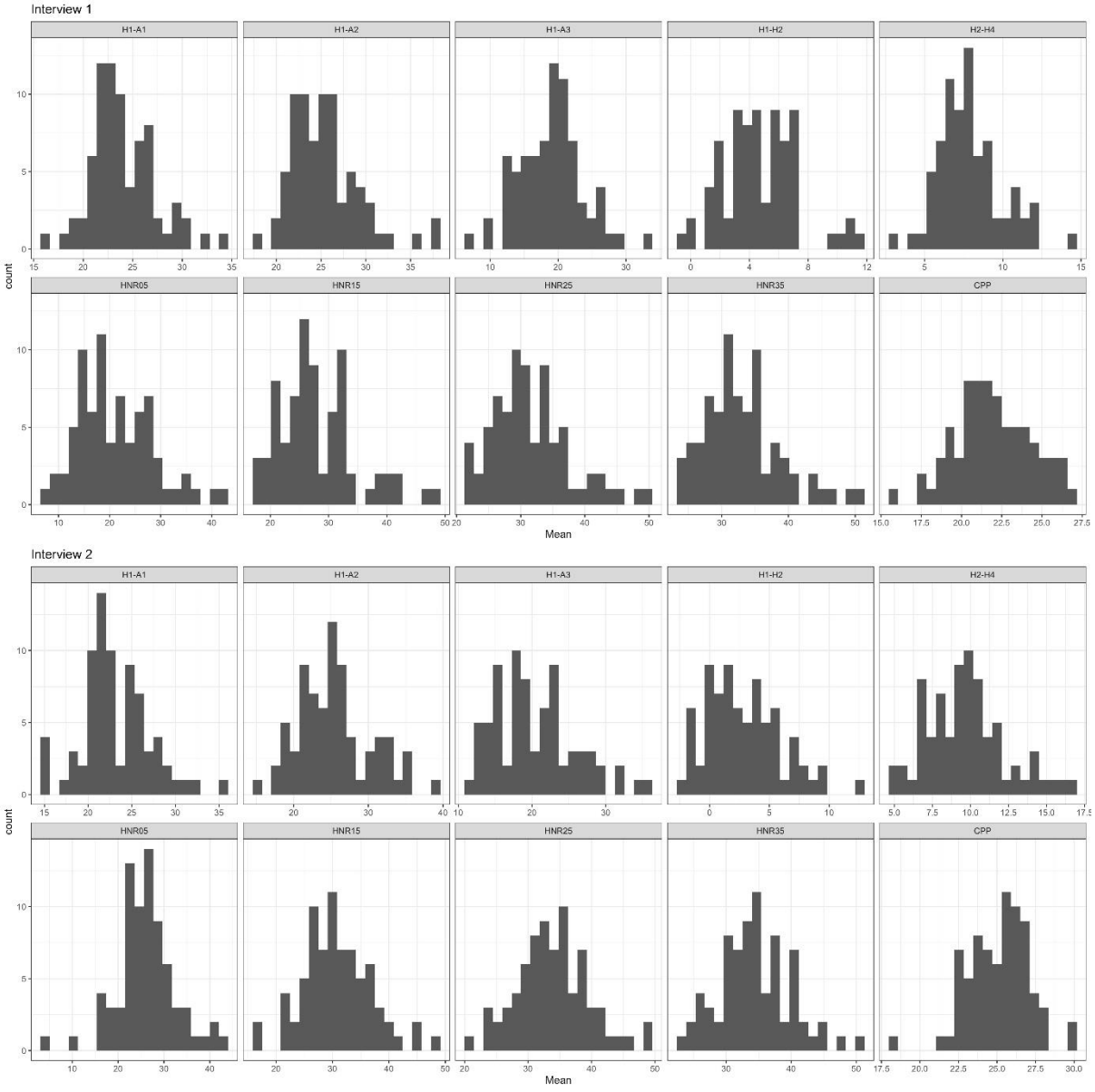
Place of origin of the speakers in the present study

	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

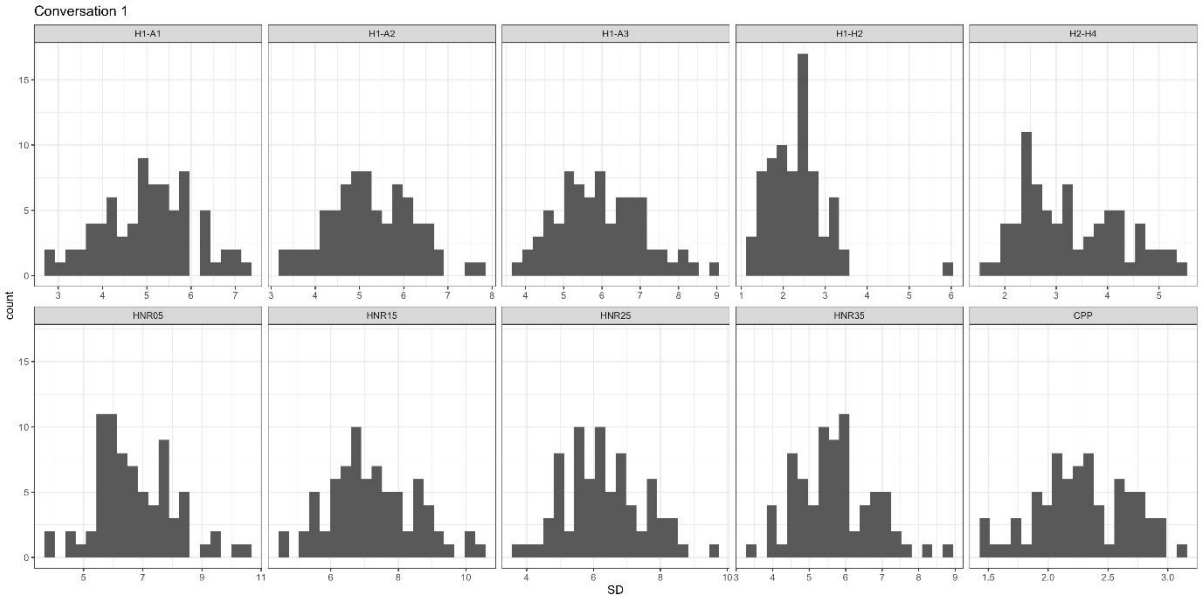
Mean values of individual VQ parameters across 75 speakers in different recordings (CNV1, CNV2, INT1, INT2)

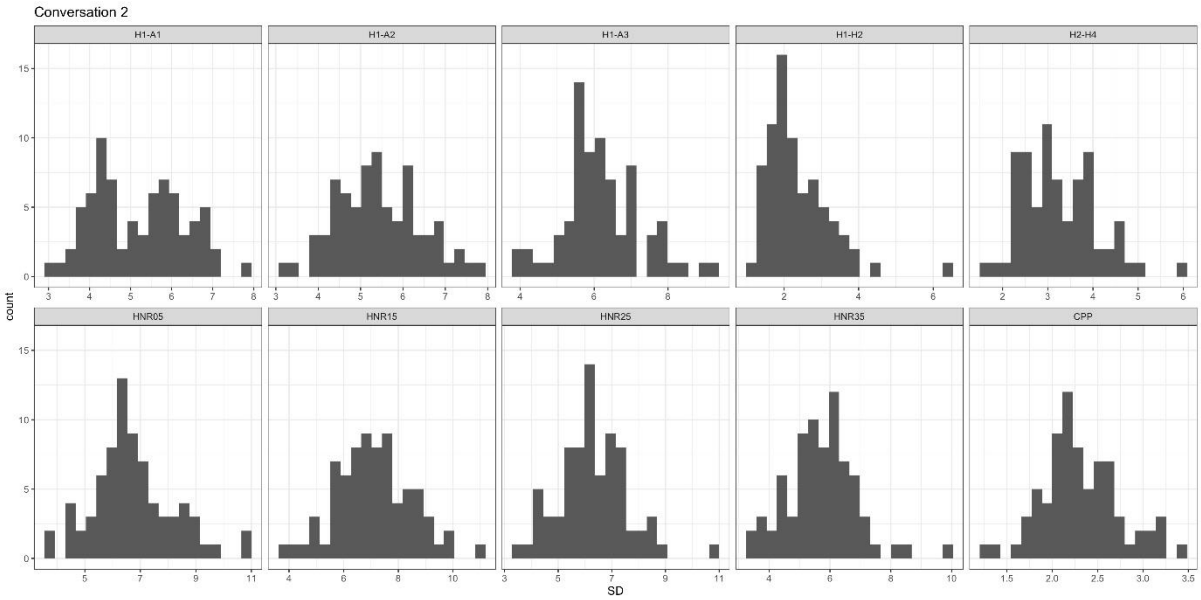


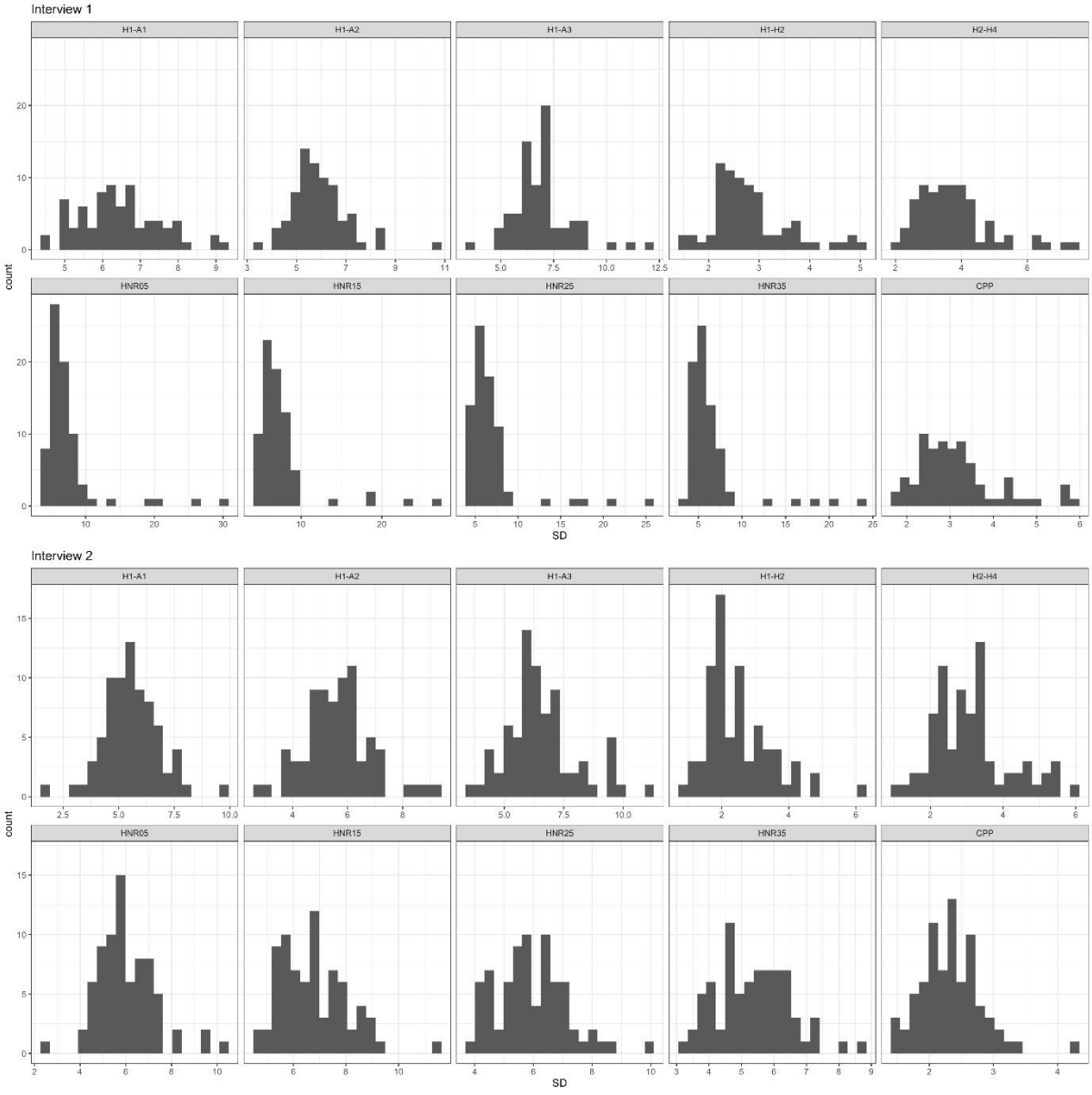

Interview 1
H1-A1
H1-A2
H1-A3
H1-H2
H2-H4
HNR05
HNR15
HNR25
HNR35
CPP
count
Mean
Interview 2
H1-A1
H1-A2
H1-A3
H1-H2
H2-H4
HNR05
HNR15
HNR25
HNR35
CPP
count
Mean

Appendix C

Standard deviations of individual VQ parameters across 75 speakers in different recordings (CNV1, CNV2, INT1, INT2)






Interview 1
H1-A1
H1-A2
H1-A3
H1-H2
H2-H4
HNR05
HNR15
HNR25
HNR35
CPP
count
SD
Interview 2
H1-A1
H1-A2
H1-A3
H1-H2
H2-H4
HNR05
HNR15
HNR25
HNR35
CPP
count
SD