Speaker discriminatory power of voice quality acoustics under forensic conditions

Ricky K. W. Chan

Speech, Language and Cognition Laboratory, School of English, University of Hong Kong
rickykwc@hku.hk

ABSTRACT

Voice quality has been regarded as one of the most popular and useful features for forensic speaker comparison, but relevant empirical validation is limited, let alone the acoustic aspects of voice quality. This study assesses the evidential strength of spectral tilt and additive noise parameters under the likelihood-ratio framework. Speech data of 75 male speakers aged 18-45 were obtained from a forensically-oriented database of Australian English speakers in Sydney/New South Wales. Results show that, contrary to previous findings, spectral tilt and additive noise parameters generally carry limited speaker-discriminatory power, especially when speech style mismatch and/or non-contemporaneous recordings are involved. Implications for forensic speaker comparison are discussed.

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

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 (e.g. hoax call, ransom demand, threatening message), and the known sample typically involves the voice of a suspect captured during a police interview [1]. 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. One of the main goals in FVC research is to empirically test the speaker-discriminatory power of speech features under conditions that are typically found in forensic casework. This paper focuses on voice quality.

'Voice quality' generally refers to the quasi-permanent characteristics of one's speech running through all the sounds from the speaker as a result of various laryngeal and supralaryngeal settings [2]. Voice quality has been regarded as one of the most useful and popular features for FVC casework [1,3]. However, empirical tests of the evidential strength of voice quality features are surprisingly limited, not to mention the acoustic aspects of voice quality. A recent study [4] tested the evidential strength of harmonic and inharmonic components of laryngeal voice quality (VQ hereafter) acoustic parameters using contemporaneous speech data of 97 Southern British English male speakers. They found that the combination of these parameters could perform very well in distinguishing speakers, and telephone or mobile phone transmission only led to a small decline in performance. Also, the addition of voice quality information was found to improve MFCCs-based system performance, especially when transmission quality degraded. The present study extends these promising findings and explores how speech style mismatch and the use of non-contemporaneous recordings may affect the performance these parameters using Australian English speech data.

2. METHODS

2.1. Corpus

A forensically-oriented database of 552 Australian English speakers (332 females and 231 males at the time of writing) [5] was selected. Each speaker was recorded on one to three or more occasions based on the protocol proposed in [6]. 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.

75 speakers were chosen from the database based on the following considerations: age, gender, regional background, and the availability of non-contemporaneous recordings. Besides, only speakers who were recorded on more than one occasion (separated by at least a one-week interval) were chosen given the importance of using non-contemporaneous recordings in the evaluation of FVC systems [6]. 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. 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 [6].

2.2. Feature extraction and parameterization

Vowel-only portions of the recordings were manually segmented and labelled in Praat. Approximately 33 seconds of net vocalic material was extracted per speaker per recording. VQ parameters reported in [4] were selected so that the findings will be comparable. They were extracted using VoiceSauce [7] 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. 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 combined as ‘spectral tilt’ for further analysis.

- *Cepstral peak prominence (CPP)*. a measure of cepstral peak amplitude normalized for overall amplitude. 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. The HNR was extracted over 0-500Hz, 0-1500Hz, 0-2500Hz, and 0-3500Hz respectively, resulting in four separate measures.

These five measures were combined as the ‘additive noise’ for further analysis. 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 [8].

2.3. Statistical analysis

To assess the evidential strength of voice quality parameters, the multivariate kernel-density (MVKD) formula [9] was used for same-speaker and different-speaker comparisons. Calibrations were conducted using logistic regression. 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 100 times with different 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 [10]. To test the effects of speech style mismatch and 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 log-LR cost (C_{llr}). The lower the C_{llr} value, the better the system performance. C_{llr} values close to or greater than 1 imply limited evidential value of the input parameter(s). Due to space constraints, equal error rate values will be presented in the conference.

3. RESULTS & DISCUSSION

CNV1 vs. CNV2: C_{llr}				
	Min	**Max**	**Mean**	**SD**
Spectral tilt	0.61	1.07	0.77	0.08
Additive noise	0.52	0.91	0.67	0.08
Spectral tilt + Additive noise	0.85	1.08	0.93	0.04

CNV1 vs. INT1: C_{llr}				
	Min	**Max**	**Mean**	**SD**
Spectral tilt	0.91	1.04	0.96	0.02
Additive noise	0.84	1.05	0.92	0.04
Spectral tilt + Additive noise	0.85	1.02	0.93	0.04

CNV1 vs. INT2: C_{llr}				
	Min	**Max**	**Mean**	**SD**
Spectral tilt	0.91	1.05	0.97	0.03
Additive noise	0.76	1.01	0.88	0.05
Spectral tilt + Additive noise	0.87	1.02	0.93	0.03

Tables 1 to 3: statistics of C_{llr} values across 100 replications with VQ parameters as input in CNV1 vs. CNV2, CNV1 vs. INT1, and CNV1 vs. INT2 respectively.

Tables 1 to 3 show the descriptive statistics of C_{llr} values based on the combined spectral tilt measures (5 parameters), the combined additive noise measures (5 parameters), and spectral tilt + additive noise (10 parameters) for the three comparisons. In general, all the input parameters yielded a rather small standard deviation in C_{llr} value (less than 0.1) across the 100 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. With non-contemporaneous recordings and the same speech style, both spectral tilt and additive noise measures yielded good results, with the best replications returning C_{llr} values of 0.61 and 0.52 correspondingly. Surprisingly, using spectral tilt + additive noise as input led to worse system performance, with the lowest C_{llr} being 0.85.

CNV1 vs. CNV2. With contemporaneous recordings and mismatch in speech style, the spectral tilt and the additive noise measures only performed slightly worse, with the best replications producing C_{llr} values of 0.91 and 0.84 correspondingly. Using both spectral tile and additive noise as input did not seem to improve system performance (best performance: C_{llr} = 0.85).

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. The spectral tilt measures did not appear to provide much speaker-discriminator information, with the best replication having a C_{llr} of 0.91. On the other hand, the additive noise measures performed better and yielded a C_{llr} of 0.76 in the best replication. However, adding spectral tilt measures to additive-noise-based system led to poor performance— a C_{llr} of 0.87 in the best replication.

4. DISCUSSION

The present study sought to determine the evidential strength of VQ acoustic parameters under the LR framework, and tested the effects of speech style mismatch and the use of non-contemporaneous recordings on system performance. In general, the additive noise measures performed slightly better than the spectral tilt measures. This suggests that additive noise measures carry more speaker-specific information and higher evidential strength for FVC. However, combining spectral tilt and additive noise measures led to worse system performances, suggesting that they may provide overlapping or even conflicting information for discriminating speakers. While it has been found that the harmonic components may interact with the inharmonic components (noise) in various frequency bands in the expression of VQ-based phonological contrasts in some languages [11]), VQ is not known to be used for phonation contrasts in Australian English. The interaction between the harmonic and inharmonic components of VQ in speaker characterization under forensically-relevant conditions deserves further investigation.

[4] used the same parameters but only with contemporaneous speech data that involved speech style mismatch (conversation vs. interview; similar to CNV1 vs. INT1 in this study). They found that the spectral tilt and the 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 this may be attributable to the methodological differences between the two studies such as the number of speakers (75 vs. 97), length of speech material per recording (approximately 33s vs. 60s), the mathematical models used for speaker modelling (MVKD vs. the Gaussian Mixture Model-Universal Background Model (GMM-UBM)), and the variety of English (Australian English vs. standard southern British English) involved. The last of these suggests that the evidential strength of VQ parameters may be language/variety-specific and cautions should be exercised when generalising the results to other languages.

In addition, the present study involves two more sets of comparison to determine the effects of speech style mismatch and non-contemporaneous recordings on the speaker-discriminatory power of VQ acoustic parameters. VQ parameters performed relatively well in CNV1 vs. CNV2 (same speech style, non-contemporaneous recordings). However, results are much less promising when speech style mismatch is involved (CNV1 vs. INT1 or INT2). The results of CNV1 vs. INT1 and CNV1 vs. INT2 are similar. These findings suggest that speech style mismatch have more detrimental effects than non-contemporaneous recordings in this study, and that the involvement of non-contemporaneous recordings does not have clear effects on system performance. This highlights 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. 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 [6], but these affect system performances to a lesser extent. Future research may investigate how the performance of VQ parameters may be affected by other speech styles typically found in forensic casework (e.g. speech under various emotional states), and with recordings separate by different time periods. With both speech style mismatch and non-contemporaneous recordings (i.e. CNV1 vs. INT2) as in typical forensic casework, all but the additive noise measure yielded high C_{llr} values, suggesting that the VQ parameters examined in this study may 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 [1,3]. Nonetheless, the present study only analysed 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. [1] note that in typical voice quality analysis using a version of the Laver VPA scheme, around 38 speech features and vocal tract settings may be analysed auditorily. 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, and such endeavour 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.

In the current study, the evidential strength of VQ parameters were assessed using multivariate kernel-density (MVKD) formula. Reanalysis of the data based on GMM-UBM is in progress and the results will be reported in the future.

5. CONCLUSION

Given the importance of empirically validating speech features to be used in FVC casework, the present study investigated the evidential strength of spectral tilt and additive noise parameters under the LR framework. It was found that these laryngeal voice quality parameters generally offer limited speaker-discriminatory value, particularly when 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.

6. REFERENCES

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

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

[3] Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech Language and the Law*, *18*(2), 293-307.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series c (Applied Statistics)*, *53*(1), 109–122.

[10] 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.

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