FORENSIC VOICE COMPARISON USING LONG-TERM ACOUSTIC MEASURES OF LARYNGEAL VOICE QUALITY

Vincent Hughes[1], Amanda Cardoso[3], Philip Harrison[1,2], Paul Foulkes[1], Peter French[1,2], Amelia J. Gully[1]

[1]Department of Language and Linguistic Science, University of York, UK
[2]J P French Associates, York, UK
[3]Department of Linguistics, University of British Columbia, Canada
{vincent.hughes|philip.harrison|paul.foulkes|peter.french|amelia.gully}@york.ac.uk

ABSTRACT

Voice quality (VQ) is reported by forensic analysts to be a useful variable in voice comparison casework. Despite this, very little research has assessed the efficacy of VQ as a speaker discriminant. This paper employs semi-automatic methods to test the performance of a forensic voice comparison system using long-term acoustic measures of laryngeal VQ. Fundamental frequency, cepstral peak prominence, harmonics-to-noise ratios and a range of spectral tilt measures were extracted from vowel-only samples of studio, landline telephone, and mobile telephone recordings. Using likelihood ratio-based testing, the VQ features produced promising results. The high quality condition produced EERs as low as 5.8% and C_{llr}s as low as 0.26, although performance was degraded in the telephone conditions. When fused with a baseline MFCC system, results were mixed, with VQ improving performance only for some configurations of speakers.

Keywords: Voice quality, acoustics, forensic voice comparison, automatic speaker recognition

1. INTRODUCTION

Voice quality (VQ) is the long-term, quasi-permanent 'timbre' of the voice which can be decomposed into a number of supralaryngeal and laryngeal settings [20,28]. The binary categorisation of settings reflects the assumed independence of the two in traditional source-filter theory [5]. VQ is often analysed using descriptive systems, such as the Vocal Profile Analysis scheme [21], which rely on impressionistic classification of articulatory settings (see [25]). VQ can also be analysed acoustically. Supralaryngeal settings can be indirectly analysed using vowel formant distributions (e.g. if a speaker has a habitually fronted tongue body, this should be reflected in a high overall F2) [6]. A range of measures has been proposed to capture laryngeal settings that, amongst other things, rely on the relationships between the amplitudes of different harmonics and cycle-to-cycle variation in fundamental frequency (F0) and amplitude. A major issue for the study of VQ is understanding the relationship between articulation, acoustics, and auditory perception [17,19].

VQ is often analysed in forensic voice comparison cases, where an expert is asked to compare the speech patterns of a known suspect and unknown offender in the context of a legal case. A survey of practitioners found that VQ is generally considered to be the most useful speaker discriminant [8]. However, very few studies have been conducted to test this claim. Those that have typically analysed acoustic measures of laryngeal VQ. [4] used the GLOTTEX software to extract a large number of voice source features from recordings of 60 female standard Mandarin speakers. They compared the performance of VQ features against a Mel-frequency cepstral coefficient (MFCC)-based automatic speaker recognition (ASR) system. The ASR system consistently outperformed VQ, and there was no improvement in ASR performance when combined with VQ features.

This finding is not predicted by the theoretical decoupling of source and filter in computing cepstral coefficients [15]. In principle, MFCC-based systems should only capture information about the supralaryngeal vocal tract and there is some evidence support this. [9] performed a phonetic analysis of the false hits produced by an iVector-based ASR. They found that it was possible to distinguish these pairs auditorily, and that phonation (in particular creak) was the most useful diagnostic. Similarly, the errors in [11] were easily resolved on the basis of laryngeal VQ. Given these findings, the results in [4] may reflect methodological decisions such as only extracting measures from the segment /n/ when used as a filler, and the small number of tokens for some speakers, rather than the inherent value of VQ as a speaker discriminant. However, they may also reflect the indirect relationships between acoustics and auditory judgements, with the two approaches capturing different speaker-specific information.

The present study performs systematic speaker discrimination testing using laryngeal VQ features. The results are compared and combined with MFCC-based ASR systems to assess the extent to which they capture complementary speaker-specific information.

2. METHOD

For a detailed overview of the recordings and processing used in this study see [12,13]. The key details are outlined briefly here. 97 male speakers of standard southern British English from the DyViS corpus were analysed [22]. For each speaker, two recordings were available, constituting nominal suspect (Task1) and offender (Task2) samples. From each recording, 60 seconds of vowel material was automatically extracted. The vowel-only samples were available in four channel conditions: high quality studio samples (HQ), landline telephone samples (TEL), and two mobile phone samples with high (12.2kb/s; MOB_{HQ}) and low (4.75kb/s; MOB_{LQ}) bit rate. These conditions are commonly found in forensic casework.

2.1. Feature extraction

The vowel-only samples were divided into a series of 20ms frames with 10ms overlap. VQ and MFCC feature vectors were extracted from each frame.

2.1.1. Laryngeal voice quality (VQ)

VQ measures were extracted using VoiceSauce [26]:

- **Cepstral peak prominence (CPP):** the normalised peak of the pitch period within the real cepstral, quefrency domain [10]
- **Harmonics-to-noise ratios (HNR):** comparing the energy of harmonics with the noise floor in the cepstral domain using the algorithm in [3], calculated over four frequency ranges: 0-500Hz, 0-1500Hz, 0-2500Hz, and 0-3500Hz
- **H1-A1; H1-A2; H1-A3:** the amplitude of the first harmonic (F0) relative to the amplitude of the harmonic closest to the first (A1), second (A2), and third (A3) formants
- **H1-H2; H2-H4:** the amplitude of the first harmonic relative to the amplitude of the second harmonic, and the amplitude of the second harmonic relative to the amplitude of the fourth harmonic.

F0 was also analysed using the straight algorithm [16] with the range set from 75Hz to 200Hz. To compare across different vowels, the spectral tilt measures were corrected using F0 and formant tracking (see [14]). The harmonics were estimated using the same procedure as for F0 extraction. Formants were estimated using the Snack Toolkit [27] tracking five formants within a range of 0 to 5000Hz (LPC order: 12, pre-emphasis: 0.96).

The variables above were chosen for analysis on the basis of their established link with auditory percepts of VQ and their relatively extensive use in the phonetics literature. CPP and HNR are additive noise measures relating to harmonic structure that, in principle, capture differences between breathy and modal voice. Breathy voice should produce a less prominent cepstral peak and lower HNR. Breathy and creaky voice are also claimed to differ in terms of their energy distribution across the spectrum, with high spectral tilt associated with breathy voice and low spectral tilt associated with creaky voice [7,18].

2.1.2 Mel-frequency cepstral coefficients (MFCCs)

From each frame, 12 MFCCs, delta (Δ) and delta-delta (ΔΔ) coefficients were extracted using the *rastamat* toolbox [23]. For the HQ samples, MFCCs were extracted within a 0 to 4000Hz range, while for the telephone and mobile samples, extraction was performed within a 300 to 3400Hz range

2.2. Scoring, calibration, fusion and evaluation

Channel was matched across the suspect and offender samples in all of the testing here, producing four conditions: HQ-HQ, TEL-TEL, MOB_{HQ}-MOB_{HQ}, and MOB_{LQ}-MOB_{LQ}. For each condition, a range of input features were analysed to assess comparative performance. Testing was initially conducted using a combination of all VQ measures. The additive noise (CPP and HNR) and spectral tilt measures ($H1\text{-}A_n$, $H_n\text{-}H_m$) were also analysed separately. The MFCC systems were tested using all of the available features.

The 97 speakers were randomly assigned to development (32 speakers), test (33 speakers) and reference (32 speakers) sets. For each set of input features, same- (SS) and different-speaker (DS) GMM-UBM [24] scores were computed for the development and test speakers using the reference speakers to assess typicality. GMMs were fitted with 512 Gaussians (for both VQ and MFCCs, based on pre-testing). The test scores were then calibrated using logistic regression coefficients [2] derived from the development scores. This produced calibrated log likelihood ratios (LLRs), which were used to assess system performance. To account for the variability in output as a function of the specific speakers used in each set, testing was replicated 20 times using random configurations of development, test, and reference speakers [29].

The output of the MFCC systems was also combined with phonatory VQ to assess whether the addition of the latter information improved performance over MFCCs in isolation. This was done using logistic regression fusion [2]; a procedure for calibrating and combining scores from multiple systems that accounts for correlations between scores.

As in 2.2, the development scores were used to generate fusion coefficients which were applied to the test scores to produce calibrated LLRs for each replication.

The performance of the VQ and MFCC systems, both separately and in combination, was analysed using equal error rate (EER) and the log LR cost function (C_{llr}; [1]). The effect of combining VQ and MFCCs was assessed by calculating the percentage difference in EER and C_{llr} between the MFCC-only (baseline) system and the fused system.

3. RESULTS

3.1. Voice quality systems

Table 1 displays mean, minimum, and maximum EER and C_{llr} values across the 20 replications within each of the four channel conditions using all of the VQ features (F0, additive noise, spectral tilt) as input. Optimal performance was found in the HQ condition, with the best performing replication producing an EER of 5.8% and a C_{llr} of 0.26. As might be predicted, performance degrades with telephone and mobile samples. However, the extent of the decrease in performance is relatively small (equivalent to an EER difference of 3-4% EER and a C_{llr} difference of 0.10). This suggests that the acoustic measures of VQ tested here are relatively robust to the channel variation commonly found in forensic casework.

Table 1: Mean, minimum, and maximum EER and C_{llr} values across the 20 replications using all VQ features as input.

	Mean		Min		Max	
	EER	**C_{llr}**	**EER**	**C_{llr}**	**EER**	**C_{llr}**
HQ	9.6	0.39	5.8	0.26	12.2	0.63
TEL	13.2	0.49	6.1	0.33	18.2	0.59
MOB_{HQ}	13.4	0.49	6.1	0.32	18.8	0.64
MOB_{LQ}	13.3	0.51	6.1	0.33	18.0	0.76

Table 2 displays the performance of the additive noise and spectral tilt measures (F0 is not included here as the focus is on laryngeal VQ). Across all channel conditions, the spectral tilt measures outperformed the additive noise measures. Indeed, in the MOB_{LQ} condition, the spectral tilt measures performed almost as well as all of the VQ features in combination. This suggests that spectral tilt encodes considerable speaker-specific information and accounts for a large proportion of the speaker discriminatory power in the systems in Table 1. The consistency in performance across conditions also indicates that spectral tilt is relatively robust to channel variation, in a way that additive noise measures are not.

Table 2: Mean EER and C_{llr} values across the 20 replications using the additive noise and spectral tilt features as input.

	Additive noise		Spectral tilt	
	EER	**C_{llr}**	**EER**	**C_{llr}**
HQ	17.6	0.61	13.1	0.54
TEL	19.8	0.69	15.1	0.59
MOB_{HQ}	20.0	0.91	15.6	0.59
MOB_{LQ}	20.7	0.82	13.9	0.54

3.2. Fusion of systems

Figure 1 shows the baseline MFCC-only and fused VQ and MFCC performance for the 20 replications in each of the conditions. For the sake of space, the C_{llr} is not reported here, but is available at [https://vincehughes.files.wordpress.com/2019/03/vqasrperformance.pdf]. In five of the 20 replications in the HQ condition, the addition of VQ information improved EER. For two of these replications, the addition of VQ produced a system which successfully discriminated between all SS and DS pairs (EER = 0%). In 10 replications no improvement was found, while for five replications the EER of the fused system was substantially worse.

However, as transmission quality degraded, the contribution of VQ to system performance became more impressive. The value of VQ was most notable in the MOB_{LQ} condition, where the addition of VQ improved EER for 16 of the 20 replications. For these 16 replications, the average decrease in EER was 70%. The largest decrease was 97% for a replication that produced an EER of 2.85% using only MFCCs and 0.09% when fused with VQ. In the other four replications EER was worse for the fused system than for the MFCC-only system. In one case the addition of VQ shifted the baseline EER of 0.38% to 2.37%.

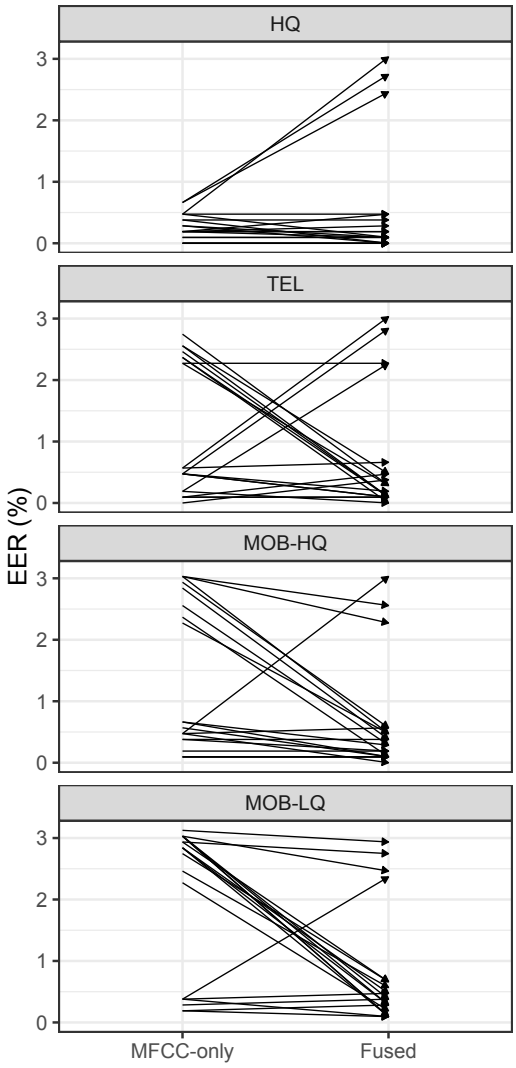
4. DISCUSSION

The results in 3.1 reveal that long-term acoustic measures of laryngeal VQ extracted from vowels capture considerable speaker-specific information. In optimal conditions, and with the right configuration of speakers, EERs of around 6% are possible. These results are extremely promising when assessed relative to other linguistic variables. Comparison with [13] shows that VQ outperforms formants on the same vowel-only material.

Performance degrades when using telephone and mobile samples. However, the magnitude of the effect was relatively small, and impressive performance was achieved even with low bit-rate mobile samples. This suggests that acoustic VQ measures are relatively robust to channel variation. This is extremely important for forensic voice comparison casework given the range of potential recordings analysed. As

shown in Table 2, the spectral tilt measures are largely responsible for the good overall speaker-discriminatory performance of VQ and its robustness to channel variation.

Figure 1: Baseline MFCC-only (left) and fused (MFCC and VQ; right) EERs across the 20 replications in each of the four channel conditions.



The results are all the more impressive in light of the automatic nature of the analysis. In extracting the VQ data, no speaker-specific adjustments were made to settings (see [13] for the effect of settings on formant extraction for forensics) and no post-processing was applied to deal with potential measurement errors. Further, a limited range of acoustic measures was used in this study, compared with the extremely large number of potential measures of laryngeal VQ available (see e.g. [4]).

While the performance of laryngeal VQ in isolation is encouraging, perhaps the most impressive results are those based on the combination of this and MFCCs. In the HQ condition, fused performance was the same as, if not worse than, the MFCCs in isolation. This is likely to be due to a ceiling effect, with all 20 replications producing EERs of less than 1% when using only MFCCs as input, leaving very little room for improvement in performance. However, as quality degraded, much more substantial improvement in performance was found when combining MFCCs with VQ. This suggests that acoustic measures of laryngeal VQ do capture complementary speaker-specific information to that captured by MFCCs, in line with theoretical predictions about the separation of source and filter in MFCC extraction. These results indicate that, in addition to MFCCs, it may be advantageous for ASR systems to also extract measures of phonation.

Throughout this study, analysis has been conducted at the system level, considering overall speaker-discriminatory performance. However, the variability across the 20 replications (regardless of the input features) shows that some speakers are easier to separate than others. This means that overall system-level performance, and the potential improvement in performance due to VQ, is dependent on the makeup of the development, test, and reference sets. There are two implications of this. First, it is important to test systems with different configurations of speakers to assess variability in performance (as discussed in [29]). Second, caution should be exercised when generalising about the speaker-discriminatory power of features or combinations of features based on overall system performance (such as in [4]). While our results are useful in a general sense, they tell us little about specific cases. For instance, knowing that laryngeal VQ generally improves ASR performance in mobile conditions does not necessarily mean it will be useful for the specific suspect and offender voices in a given case. Rather, as we highlight in [12], more research in forensic voice comparison should focus on understanding the behaviour of individual speakers within systems, to try to better understand the conclusions we arrive at in casework.

5. CONCLUSIONS

This study has examined the speaker discriminatory value of long-term acoustic measures of laryngeal VQ. The results confirm analysts' intuitions that VQ is an extremely useful variable in forensic voice comparison cases. VQ has also been shown to be capable of improving ASR performance, especially when channel is degraded.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

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

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

[3] de Krom, G. 1993. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *J. Sp. Lang. Hear. Research* 36.2, 254–266.

[4] 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.

[5] Fant, G. 1960. *Acoustic Theory of Speech Production*. The Hague: Mouton.

[6] French, J.P., Foulkes, P., Harrison, P., Hughes, V., Stevens, L. 2015. The vocal tract as a biometric: output measures, interrelationships and efficacy. *Proc. 18th ICPhS* Glasgow.

[7] Garellek, M. 2017. The phonetics of voice. *Routledge Handbook of Phonetics*.

[8] Gold, E., French, J.P. 2011. International practices in forensic speaker comparison. *IJSLL* 18.2, 293–307.

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

[10] Hillenbrand, J., Cleveland, R.A., Erickson, R.L. 1994. Acoustic correlates of breathy vocal quality. *J. Sp. Lang. Hear. Research* 38.6, 769–778.

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

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

[13] Hughes, V., Harrison, P., Foulkes, P., French, J.P., Gully, A. 2019. Effects of formant analysis settings and channel mismatch on semi-automatic forensic voice comparison. Submitted to *ICPhS*, Melbourne.

[14] Iseli, M., Shue, Y.-L., Alwan, A. 2007. Age, sex, and vowel dependencies of acoustic measures related to the voice source. *J. Acoust. Soc. Am.* 121.4, 2283–2295.

[15] Jurafsky, D., Martin, J.H. 2009. *Speech and Language Processsing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics* (2nd ed). Prentice-Hall.

[16] Kawahara, H., de Cheveigné, A., Patterson, R.D. 1998. An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite. *Proc. 5th Int. Conf. on Spoken Language Processing*, Sydney.

[17] Keating, P., Garellek, M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proc. 18th ICPhS* Glasgow.

[18] Klug, K., Kirchhübel, C., French, J.P., Foulkes, P. 2018. Do the acoustics support the perception? The example of breathy voice. *Paper presented at IAFPA conference*, Huddersfield.

[19] Kreiman, J., Shue, Y.-L. 2010. Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *J. Acoust. Soc. Am.* 132.4, 2625–2632.

[20] Laver, J. 1980. *The Phonetic Description of Voice Quality*. Cambridge: CUP.

[21] Laver, J., Wirz, S., Mackenzie Beck, J., Hiller, S. 1981. A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress*. 14, 139–155.

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

[23] Rastamat Toolbox for MATLAB. http://labrosa.ee.columbia.edu/matlabl/rastamat

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

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

[26] Shue, Y.-L. 2010. *The Voice Source in Speech Production: Data, Analysis and Models*. UCLA dissertation.

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

[28] Trask, R.L. 1996. *A Dictionary of Phonetics and Phonology*. London: Routledge.

[29] Wang, B., Hughes, V., Foulkes, P. 2019. The effect of speaker sampling to system stability in likelihood ratio-based forensic voice comparison. Submitted to *ICPhS*, Melbourne.