The vocal tract as a biometric: output measures, interrelationships, and efficacy

Peter French, Paul Foulkes, Philip Harrison, Vincent Hughes, Eugenia San Segundo and Louisa Stevens

University of York and J P French Associates
{peter.french| paul.foulkes|philip.harrison|vincent.hughes|eugenia.sansegundo|lcc108}@york.ac.uk

ABSTRACT

This paper explores methods for characterising individual voices using different vocal tract output measures. Mel frequency cepstral coefficients (MFCCs), long-term formant distributions (LTFDs) and scores based on vocal profile analysis (VPA) of long-term supralaryngeal settings were extracted from the same corpus of recordings. Distances between speakers were calculated and used to test the interrelationships between the three output measures. Strong correlations were found between the MFCC and LTFD distances, while considerably weaker correlations were found between the acoustic measures and the VPA-based distances. This suggests that while the two measures of acoustic output provide similar information, the auditory VPA offers different information relevant for voice characterisation. In a forensic context this finding is important since it suggests that it may be possible to complement acoustic analysis with VPA to improve system performance.

Keywords: individual speaker characterisation, forensic speaker comparison, vocal tract output measures.

1. INTRODUCTION

It is commonly assumed that individual voices are unique, shaped by a combination of the talker's biology and socialisation. The speaker's anatomy and physiology provide a basic architecture that is styled with acquired patterns of behaviour from the regional and social groups in which the speaker learns his language(s). The research presented here arises from a project on individual speaker characterisation entitled *Voice and Identity – Source, Filter, Biometric*. The broad aim of the project is to contribute to our understanding of the parameters on which speakers may systematically be differentiated from one another, and to propose ways in which analyses of those parameters may be combined into a battery of tests for forensic speaker comparison cases. The first phase of the work (reported here) concerns the investigation of the vocal tract as a potential source of information for individual speaker discrimination.

The issue of individual speaker variation is of central concern in the forensic domain. In a typical forensic scenario, an expert is instructed to compare the voices in recordings of a known suspect and an unknown offender to aid the court in determining whether the suspect and offender are the same or different talker(s). Within the field of speech technology, automatic speaker recognition (ASR) systems attempt to capture the essential properties that distinguish one voice from another. However, the complexities of forensic cases cause difficulties for ASR systems (e.g. through channel mismatch, background noise). Although forensic casework worldwide is increasingly drawing on ASR, there is to our knowledge no jurisdiction in which speaker comparisons are undertaken purely by ASR unaided by human assessment to some extent [7,9]. Indeed, in many jurisdictions ASR is used very little, and the task is still largely handled by phoneticians applying analytic methods drawn from phonetics and linguistics [5]. Underlying many of the linguistic-phonetic and ASR analyses of samples is the assumption that the vocal tract is a biometric: that is, the assumed uniqueness of the physical vocal tract can be modelled to separate individual voices from one another. While the assumption of vocal tract uniqueness is almost certainly sound, it is not possible to examine the vocal tract directly. Speech scientists analyse different measures of the *acoustic output* of the vocal tract – analogous to modelling the barrel of a gun from the sound of its shot. However, there has been little if any comparative assessment of the relative contribution of different output measures in speaker discrimination.

In this paper we explore the value of vocal tract output in characterising individual voices. We compare the performance of different output measures, and their interrelationships. We focus on:

- MFCCs (mel frequency cepstral coefficients) – commonly used features in ASR systems;
- LTFDs (long-term formant distributions) – a global (i.e. non-segmental) analysis of the distributions of formants across a recording, which provides information about the vowel system and the vowel space, increasingly used in linguistic-phonetic forensic research [8];
- VPA (vocal profile analysis) – an auditory-based analysis of long term vocal settings and voice quality, developed largely for speech pathology analysis but also commonly used in mainstream phonetics [13].

2. MATERIALS AND METHOD

2.1. Materials

Data for analysis were drawn from the DyViS corpus [14]. DyViS contains recordings of 100 male speakers of Standard Southern British English (SSBE), aged 18-25. We used data from Task 2: spontaneous speech elicited via a telephone conversation relating to a mock crime. For this study, high quality recordings of the target speaker at the near end of the telephone line were used (i.e. the signal was not transmitted via the telephone).

2.2. Measures

Three measures of vocal tract output were implemented: MFCCs, LTFDs and VPA. Each was used to calculate a distance measure and an identification score. The distance measure quantifies the degree of divergence between each pair of voice samples. The identification score assesses how well the features can be used to classify pairs of samples as ‘same speaker’ (SS) or ‘different speaker’ (DS).

2.2.1. MFCC analysis

MFCCs were extracted and processed using the commercial ASR system BATVOX (v4). Silences were removed automatically, leaving the speech-active portion of each sample. The signal was then divided into frames using a 20ms Hamming window shifted at 10ms steps, resulting in 50% overlap between adjacent frames. From each frame, a feature vector of 20 MFCCs, 20 delta and 20 delta-delta coefficients was extracted and used to build a Gaussian Mixture Model (GMM: 1024 Gaussians) for each speaker. Kullback-Leibler (KL) divergences were calculated to quantify the distance between speaker models.

2.2.2. LTFD analysis

For the LTFD analysis, the recordings were first subjected to automatic vowel segmentation using StkCV software [2]. For consistency, these vowel-only samples were reduced to 50 seconds net speech (the duration of the shortest sample after segmentation). Previous studies have also shown that LTFD models stabilise at around 50 seconds [8]. The samples were then analysed by the iCAbS formant tracker [12], logging measurements of the first four formants using a 25ms Gaussian-like window shifted at 5ms steps. The LTFDs were then fitted with a GMM (8 Gaussians) and KL distances again calculated between each speaker pair. Means (LTFMs) were also calculated for each formant.

2.2.3. VPA analysis

The voice samples were analysed using a modified version of the Edinburgh VPA [13], containing 28 supralaryngeal dimensions with seven scalar points. The sixth author undertook the assessment. Divergence between speaker pairs was quantified as the Euclidean distance over the 28 dimensions. Distance scores ranged between 0 and 9. This range was, as expected, relatively narrow. Only a small subset of dimensions received scores above 1 for any speaker, as the purpose of the VPA is to capture habitual (i.e. long-term) departures from a well-defined neutral setting. Scalar values of 4 and higher are restricted to speakers with voice/speech disorder, and were rarely if ever used in our analysis. Moreover, it is clear that VPA dimensions are not independent of one another. For instance, open and close jaw cannot occur simultaneously, and thus a score above 0 for one predicts 0 for another.

2.3. Method

2.3.1. Correlations

Two sets of correlation tests were performed to explore the interrelationships between the three measures. Overall correlations were first analysed using the distances from each speaker pair. These correlations were then explored in more detail, as a means of understanding the relationship between the auditory-based VPA and the acoustic-based LTF analysis. Correlations were tested between the LTFMs for each individual formant and individual dimensions on the VPA scheme. These were then compared with predictions based on phonetic theory.

2.3.2 Speaker discrimination

Different techniques were used to evaluate speaker discriminatory performance for the three measures. For MFCCs and LTFDs likelihood ratios (LRs) [16] were computed for each SS (100) and DS (4900) pair. For the MFCC analysis, LRs were computed using BATVOX in identification mode by dividing each 4 minute sample in half in order to create ‘suspect’ and ‘offender’ data. Testing in this way, rather than using non-contemporaneous samples, has been criticised as it risks failing to capture within-speaker variation adequately [4]. However, in this experiment our aim was to explore the performance of the methods under optimal conditions. For the LTFD analysis, LRs were computed using the GMM-UBM approach [15] (10 Gaussians per model, LTFD1~4). GMMs were generated from the first half of the data for comparison with the measurements from the third and fourth quarters.

Comparisons were performed with two sets of 50 speakers. Two UBMs were built using the data from 30 speakers not in each test set. LRs were transformed using a base-10 logarithm and used as a discriminant function whereby SS pairs generating a log LR of < 0 (‘miss’) and DS pairs generating a log LR of > 0 (‘false hit’) were classed as errors.

A different approach was used to analyse the speaker discriminatory value of the VPA data, owing to the current lack of formulae for adequately computing LRs for discrete data [1,6]. Further, only one data set was available per speaker, meaning that SS comparisons were not possible. VPAs from different speaker pairs were compared to establish the number of exact (i.e. 1:1) matches. Given that 100% agreement between raters on such a complex protocol is unlikely, a less stringent criterion was applied to establish close correspondences between pairs. Pairs differing by 2 scalar values or fewer were classified as ‘close’ matches. This was used as a discriminant function such that closely matching pairs were classed as false hits.

3. RESULTS

3.1. Interrelationships

3.1.1. Overall interrelationships

Table 1 summarises the overall correlations between the three vocal tract output measures based on distances between speakers. Table 1 reveals strong correlations between the two acoustic measures, LTFDs and MFCCs. The highest correlation coefficient (0.535) was found when comparing data from the first through third formants and the MFCCs, with a marginal decrease in *r* with the inclusion of F4.

Table 1: Correlations of overall distance scores between speakers (N pairs = 4950)

Comparison	*r*	*p*
LTFD1~4 vs. MFCC	0.49	<0.01
LTFD1~3 vs. MFCC	0.54	<0.01
LTFD1~4 vs. VPA	0.12	<0.01
MFCC vs. VPA	0.17	<0.01

The relationships between the acoustic measures and the VPA-based distances were, however, considerably weaker. Although the MFCCs account for marginally more variance in the VPA data than the LTFDs, the fact that the correlation coefficients for both comparisons are so small suggests that the speaker-specific information encoded in the VPA data is essentially orthogonal to that in the LTFDs and MFCCs. The small value for *p* in these cases is considered an artefact of the large amount of data.

3.1.2. Detailed interrelationships

This section explores the correlations in Table 1 in more detail. Table 2 shows the correlations between the LTFD distances for each formant separately with the distances from the other two methods.

Table 2: Correlations of distance scores between individual formants in LTFD analysis with MFCC and VPA methods (N pairs = 4950)

Comparison	MFCC		VPA	
	r	*p*	*r*	*p*
F1	0.265	<0.01	0.033	<0.05
F2	0.298	<0.01	0.073	<0.01
F3	0.439	<0.01	0.066	<0.01
F4	0.130	<0.01	0.130	<0.01

Moderately strong interrelationships were found between the individual formants and the MFCCs, with F3 providing the strongest relationship. Despite this, no comparison involving any individual formant provided a higher correlation coefficient than when combining formants. Further, consistent with the results in Table 1, the comparison involving F4 produced the weakest correlations.

As in Table 1, the correlations between the formant distances and the VPA distances are considerably weaker. The highest correlation coefficient is found for F4, although even this is relatively small. In terms of speaker distances, therefore, the VPA data are considered independent of the formant distributions. Specific correlations between individual formants and VPA dimensions were analysed using the raw data. For this the LTFMs were used to reflect the central tendency, rather than information about the entire distribution. Table 3 summarises the strongest correlations.

Table 3: Subset of strongest correlations between LTFMs for individual formants and VPA dimensions

LTFM	VPA dimension	*r*	*p*
F1	pharyngeal expansion	-0.239	<0.05
	pharyngeal constriction	0.224	<0.05
	raised larynx	0.373	<0.01
	lowered larynx	-0.226	<0.05
F2	fronted tongue body	0.270	<0.01
	lowered larynx	-0.226	<0.05
	tense vocal tract	0.197	<0.05
F3	tense vocal tract	0.273	<0.01
F4	pharyngeal constriction	-0.217	<0.05
	raised larynx	-0.419	<0.01

Contrary to Tables 1 and 2, a number of correlations were found when using the raw LTFMs and comparing with individual vocal settings. Many of these correlations were also predicted. For example, the auditory impression of fronted and backed tongue body is largely the property of vowels, and thus was correctly predicted to correlate with F2 as the key acoustic reflex of tongue position on the front-back plane. Raised or lowered larynx settings were predicted to correlate with F1 as the articulatory process of shifting the larynx affects the length of the vocal tract.

3.2. Speaker discrimination (identification)

Table 4 displays the performance of each of the three measures in the speaker discrimination task. Results are shown for true rejection (DS correctly classified), false acceptance (DS wrongly classified as SS), true acceptance (SS correctly classified), and false rejection (SS wrongly classified as DS).

Table 4: Speaker discrimination performance (%)

	MFCC	LTFD	VPA (exact)	VPA (close)
True rejection	97.1	97.4	99.5	87.9
False acceptance	2.9	2.6	0.5	12.1
True acceptance	100.0	94.0	-	-
False rejection	0.0	6.0	-	-

All three methods performed relatively well. The best performing system in terms of DS classification was that based on MFCCs and LTFDs with 3% of DS pairs producing log LRs of greater than 0. DS discrimination was somewhat lower for the VPA data. On SS discrimination, the MFCC system outperformed the LTFDs with all 100 SS pairs generating LRs greater than 0, compared with 94% for the LTFDs. This latter finding is in broad correspondence with previous research [3,10].

4. DISCUSSION

Evaluation of the interrelationships between the three forms of vocal tract output revealed strong correlations between the LTFD and MFCC distance scores. This suggests that the two acoustic measures provide similar information in terms of categorising individual voices. However, weaker correlations were found between the acoustic measures and the auditory-based VPA, suggesting the latter provides different types of information about the supralaryngeal vocal tract. Interestingly, while some of the predicted correlations between formants and VPA settings are borne out in our data, this is only the case when considering the LTFMs as an indicator of central tendency. Thus, analysis of the entire distribution of the LTFs provides complementary information to the LTFMs themselves.

The overall performance of each method was found to be very good, with errors of maximally 12% (false acceptance for VPA ‘close’ matches). This indicates that the vocal tract itself provides a considerable amount of useful information for characterising individual voices. Inevitably, all measures yielded errors. However, given the results of the correlation tests in §3.1., there is reason to expect that the acoustic measures produce different errors from those produced by the auditory analysis. Therefore, consistent with [9], these results indicate that there is considerable potential for improving the already impressive speaker discriminatory power of long term acoustic measures by complementing these analyses with auditory-based VPA.

5. CONCLUSION

This study has assessed the value of the vocal tract as a biometric by considering the interrelationships between different output measures and their relative speaker discriminatory power. The weak correlations between acoustic and auditory measures indicate that the different measures encode different types of speaker-specific information. In future work we will therefore consider how to improve speaker discriminatory performance, beyond the levels reported here, by complementing long-term acoustic analysis with VPA. If there is any value in combining MFCC analysis with LTFDs, statistical compensation procedures that take account of the correlations would need to be implemented in order to avoid duplication of information and the resulting overestimation of the strength of the evidence [4].

6. ACKNOWLEDGEMENTS

This research was conducted as part of the Arts & Humanities Research Council (AHRC) funded grant entitled *Voice and Identity – Source, Filter, Biometric* (AH/M003396/1). We wish to thank *Agnitio* for their support which has allowed us to conduct this research. Thanks also to Erica Gold for her contributions to pilot work for the project.

7. REFERENCES

[1] Aitken, C.G.G., Gold, E. 2013. Evidence evaluation for discrete data. *Forensic Science International*, 230(1-3), 147-155.

[2] Andre-Obrecht, R. 1988. A new statistical approach for automatic speech segmentation. *IEEE Trans. on ASSP*, 36(1).

[3] Becker, T., Jessen, M., Grigoras, C. 2008. Forensic speaker verification using formant features and Gaussian mixture models. *Proc. Interspeech 2008*, 1505-1508.

[4] Enzinger, E., Morrison, G.S. 2012.The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems. *Proc. 14th ASSTA conference.*

[5] Foulkes, P., French, P. 2012. Forensic phonetic speaker comparison. In: Solan, L., Tiersma, P. (eds), *Oxford Handbook of Language and Law*. Oxford: OUP, 557-572.

[6] Foulkes, P., French, P., Gold, E., Hughes, V., Harrison, P., Stevens, L., Aitken, C.G.G., Neocleous, T. 2013-15. Modelling features for forensic speaker comparison. British Academy/Leverhulme Trust Small Research Grant.

[7] French, P., Stevens, L. 2012. Forensic speech science. In: Jones, M., Knight, R. (eds), *The Bloomsbury Companion to Phonetics*. London: Continuum.

[8] Gold, E. 2014. *Calculating likelihood ratios in forensic speaker comparison cases using phonetic and linguistic features*. Unpublished PhD Thesis, University of York.

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

[10] Gold, E. French, P., Harrison, P. 2013. Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proc. Meeting on Acoustics 19*.

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

[12] Harrison, P., Clermont, F. 2012. The influence of LPC order on the accuracy of formants measurements across speakers. *IAFPA Conference*, Santander, Spain.

[13] Laver, J.D., Wirz, S.L., Mackenzie, J., Miller, S. 1981. A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress*, 14, 139-55.

[14] 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(2), 31-57.

[15] Reynolds, D.A., Quatieri, T.F., Dunn, R.B. 2000. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10, 19-41.

[16] Rose, P. 2002. *Forensic Speaker Identification*. London: Taylor and Francis.