Just one word: An analysis of *just* as a speaker discriminant using various acoustic measures

Ben Gibb-Reid

University of York
ben.gibb-reid@york.ac.uk

ABSTRACT

The task of forensic voice comparison (FVC) requires phonetic variables that are frequent in speech and easy to measure. The aim of this paper is to investigate the value of a frequent word in British English, *just*, as a speaker discriminant by comparing its performance across simple and complex combinations of measurements. Various likelihood ratio (LR)-based FVC systems are evaluated to assess the performance of vowel formants, segment and word durations and centre of gravity (CoG) measurements individually and in combination.

Promisingly, *just* performs better than STRUT and the vowel in *um*. Overall, including vowel durations and /s/ CoG measurements in a model offers slight improvements, but *just* F1~F3 vowel measurements alone perform well as speaker discriminants. In the case of *just*, an increased complexity of analysis does not offer a useful improvement in performance.

Keywords: forensic phonetics, acoustic phonetics, forensic voice comparison

1. INTRODUCTION

An ongoing goal of forensic speech science is to find suitable features of the voice which are highly variable between speakers but have low variation within an individual's speech [1,2]. To ensure comparison across samples of speech in forensic voice comparison (FVC), it is also necessary that these features are frequent, easy to measure, and ideally resistant to disguise and robust in transmission [2]. Looking at the most frequently spoken words could therefore provide a good source of potential FVC features. This study aims to assess the suitability of one such word: *just*. This is the 17th most frequent word in the spoken 2014 British National Corpus (BNC) [3], and it is increasing in usage in English when compared with the 2007 BNC (where it was the 42nd most used word). There are studies which also show increased usage in Tyneside [4] and Toronto English [5].

Although *just* has potential for the task of FVC, it is not always easy to measure. Frequent words are more likely to reduce phonetically [6], and *just* is no exception. For example, reliable formant estimates cannot be extracted from a token of *just* where the vowel is elided. As *just* is a word canonically made up of four segments /d͡ʒʌst/, there is the potential to combine various segmental and long-term acoustic measurements, with the aim of producing a more robust FVC analysis than one which relies on formants alone.

The measurements selected for analysis here are vowel formants, fricative centre of gravity (CoG), segment and word durations, and Mel-Frequency Cepstral Coefficients (MFCCs). Most of these are *single parameter* measurements, involving one segment only – for example vowel formants are only estimates of the vowel quality and do not pay attention to the whole token. However, word durations and MFCCs are comparatively *long-term* measurements and capture information from a broader range of the token. It is predicted that long-term measurements will perform better than single-parameter ones as they offer information from the entirety of a token. Due to their holistic nature, long-term measurements are also arguable easier to measure than single parameter ones as they do not rely on segmental variation to the same degree. Understanding how these measurements (and their combinations) affect the discriminatory power of *just* assesses the potential of word-based phonetics as new 'features' for the FVC toolkit. First, the performance of *just* vowel formant combinations is evaluated, and contrasted with the performance of vowel formants from *um* and STRUT. This is followed by an analysis of *just* durations and CoG measurements and a brief description of the MFCC results. The combinations of all features is also assessed and finally there is a discussion of the implications of these results for FVC.

2. METHODS

2.1. Data and token extraction

Data is taken from Task 1 in the DyViS corpus [7] – 100 male speakers of standard southern British English recorded in mock police interviews. Tokens were annotated manually and measured using Praat [8] with formant settings adjusted to look for five formants with a formant ceiling of 55kHz and a window length of 25ms. 1,019 tokens of *just* were extracted as suitable for vowel formant analysis, from

76 speakers. 92 vowel tokens were excluded due to fricative coarticulation, overlapping talk or short durations which made formant readings unreliable. In addition, 496 tokens of the vowel in *um* and 584 tokens of STRUT were extracted from the same 76 speakers for comparative analysis. STRUT is selected as the canonical vowel in *just*, and *um* is selected as it has performed well in FVC testing in the past [12]. 596 tokens of *just* from 55 speakers were also extracted as suitable for segment duration, CoG and MFCC measurements. 12 MFCCs were extracted from 20ms frames within 10ms shift across each token of *just* using a MATLAB script [9]. These results were then processed within the likelihood-ratio (LR) framework using various scripts in R, tidyverse packages were utilised to plot the results of these analyses [11,12].

2.2. Likelihood Ratios

To analyse potential FVC features, likelihood ratio (LR)-based testing is undertaken. LRs are a logical way to numerically express a conclusion comparing features of the voice and they are a broadly accepted way of presenting voice comparison evidence [13]. LR testing involves assessing the potential of a system to separate same-speaker (SS) and different-speaker (DS) pairs. In this study there are three stages of analysis which were achieved using the *fvclrr* package in R [14]. The first two (*feature-to-score* and *score-to-LR*) are achieved following previous research on potential FVC features [12]. The third stage (*replication*) follows the methods of [15]. In testing, the data is split into three subsets (test, training, and reference) before comparisons are run. Firstly, LR-like scores are calculated for the training data set against a set of reference data (*feature-to-score* stage) which assesses typicality using Multivariate Kernel Density (MKVD) [16]. Then, these test scores are calibrated using scores made from a training dataset (*score-to LR* stage). This calibration stage uses separate training data to make a logistic-regression model, which converts the LR scores to log likelihood ratios (LLRs) [17]. After this, the results were validated (*replication* stage). The first two stages were run with 25 replications – varying the arrangement of speakers across the training, test, and reference datasets. For example, one replication may have had speakers 1–25 in the training subset, speakers 26–50 in the test subset and speakers 51–76 in the background subset. Further replications would arrange the speakers in different permutations.

Two metrics are used to measure the validity of LRs in the study, the equal error rate (EER) – a measure of how many errors a system makes – and the log LR cost function (C_{llr}) [18], a measure of the severity of errors. The lower these are, the better a system is performing. For all results, the mean EER and C_{llr} across all replications is presented to ensure that variability in performance caused by sampling is accounted for. As speakers are compared against each other, a lower limit of six tokens is set to ensure no fewer than three same-speaker comparisons occurred. Any speakers who had less than six tokens which contained the required segments for analysis were excluded from that system. This excluded 24 speakers for vowel measurements and 45 speakers for segment duration, CoG and MFCC measurements.

3. RESULTS

The following section outlines the performance of each acoustic feature in various combinations. The EER and C_{llr} values for each system are used to assess them against each other in the form of averages across replications and the number of replications in which there is an improvement between systems.

3.1 Formants

Firstly, the performance of combinations of vowel formant is assessed and displayed in Figure 1.

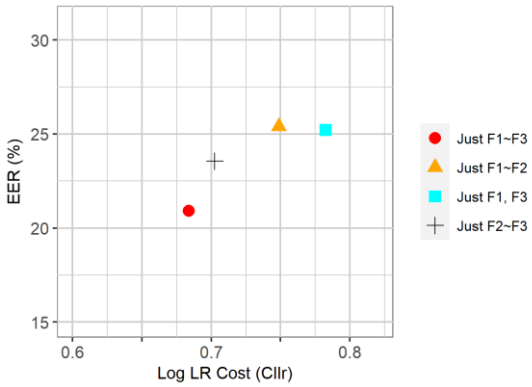

Figure 1: Mean C_{llr} plotted against mean EER comparing F1, F2 and F3 *just* vowel formant combinations across replications.

Just F1~F3 performs best overall. It has a lower mean EER (21.2%) than other systems, and a slightly lower mean C_{llr} (0.66, compared to 0.70 for F2~F3), meaning that a system relying on F1~F3 produces fewer errors of a smaller magnitude than a system based on any other combination of formants. F1~F3 has a lower C_{llr} than F2~F3 across 22 of the 25 replications and a lower EER for 21 replications, showing robustness across variable samples.

Figure 2 shows a comparison of *just* FVC performance against STRUT and *um*. Surprisingly *um* F1~F3 performs worse than *just* with a C_{llr} of 0.75 and an EER of 22.1%–more errors and of a greater magnitude. One reason for this difference in

performance could be that there were fewer tokens of *um* in the model than *just*. *Um* has a higher C_{llr} than *just* in 19 out of the 25 replications and a higher EER in 13 replications. However, these are smaller differences compared with the low performance of STRUT (0.80 C_{llr}, 30.6% EER)–a higher C_{llr} than *just* in 22 out of 25 of the replications and a higher EER in all 25 replications.

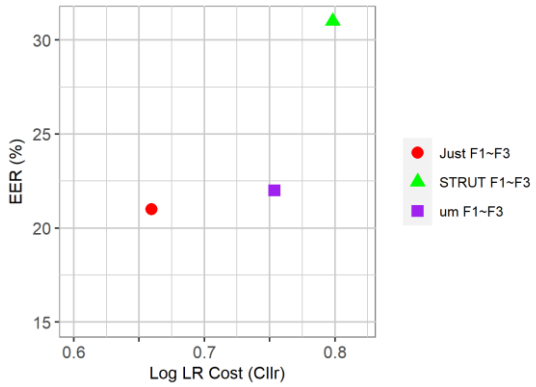

Figure 2: Mean C_{llr} plotted against mean EER comparing F1~F3 formant estimates from *just,* STRUT and the vowel in *um* across replications.

3.2 Durations

Figure 3 displays the performance of duration measures. The models are compared with a system which contains only F1~F3 vowel estimates (as there are fewer speakers, the performance of this F1~F3 system is slightly reduced against the one displayed in Figures 1 and 2).

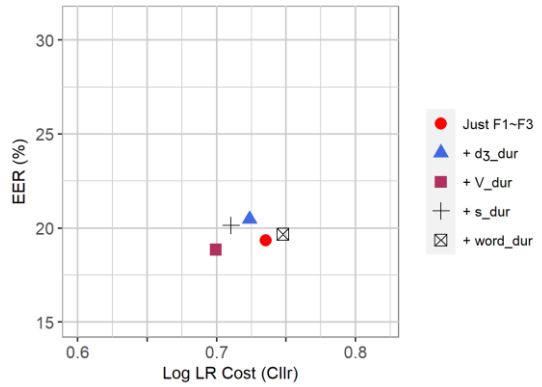

Figure 3: Mean C_{llr} plotted against mean EER comparing just segment duration measurements across replications.

There are very slight differences in performance when including segment or word durations compared with a system containing only F1~F3 (19.35% EER, 0.74 C_{llr}). Including word duration reduces model performance (a change of 0.37% EER and 0.01 C_{llr}). Including /d͡ʒ/ and /s/ duration measurements in a model leads to a very slight improvement in mean C_{llr} (by -0.02 and -0.03 respectively), with an improvement seen across 14 out of the 25 replications for /d͡ʒ/ and 17 replications for /s/. However, /d͡ʒ/ and /s/ also slightly increase EER scores (by 1.12% and 0.08% respectively). This is seen across 17 replications for /d͡ʒ/ and 13 replications for /s/. The biggest effect, though, is the improvement in performance seen when vowel duration is included. A system which contains F1~F3 and vowel duration has a C_{llr} 0.04 lower and an EER 0.5% lower than a system containing F1~F3 (this occurs in 20 out of the 25 replications for C_{llr} and 17 for EER scores). Adding vowel durations to a model therefore improves its FVC performance and it also reduces C_{llr} and EER standard deviation.

3.3 Centre of gravity

Centre of gravity (CoG) measurements were extracted from /dʒ/ and /s/. Mean performance across systems containing CoG measurements is shown in Figure 4. Three systems were tested: one containing tokens with F1~F3 vowel measurements and both CoG measurements, and two containing the vowel formant measurements and one CoG measurement.

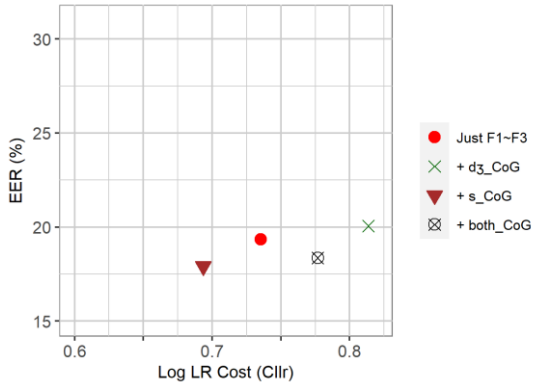

Figure 4: Mean C_{llr} plotted against mean EER comparing just centre of gravity measurements across replications.

Adding /s/ CoG measurements to a voice comparison system improved its performance. This is true for C_{llr} (a decrease of 0.04) and EER (a decrease of 1.43%) scores. Improvement is seen across 23 out of the 25 replications for C_{llr} scores and 20 replications for EER scores. Adding /dʒ/ CoG to a system reduced model performance (an increase in mean C_{llr} of 0.08, and in mean EER of 0.71%). However, an increase in C_{llr} and EER for /dʒ/ CoG is only seen across 12 out of the 25 replications. With a standard deviation of 5.6% for EER scores, /dʒ/ CoG seems to be less stable than vowel formants or /s/ CoG in terms of *just* voice comparison performance. A system containing both CoG measurements performed in-between systems containing only /s/ or /dʒ/ CoG – a slight improvement for mean EER, but an increase In mean C_{llr}. The addition of /dʒ/ CoG measurements to a system which contains /s/ CoG does not improve voice comparison performance.

3.4 MFCC measures of *just*

MFCC measurements were extracted from 726 tokens of *just* across 55 speakers. Calibrated LR voice comparison tests were run for 25 replications. The results of these suggest that MFCCs performed poorly, with a mean C_{llr} of 1.58 and a mean EER of 43.3%. C_{llr} and EER scores increase across 24 out of the 25 replications. However, this result is not a reliable reflection of the potential of MFCCs but rather a well-known statistical modelling issue, as is discussed in section 4.

3.5 Combining measures

The analysis of measurements of *just* has shown the following: for vowel formant estimates, a system containing F1~F3 performs best, and including vowel durations and /s/ CoG measurements improves model performance. A combined system made up of the best-performing single-parameter measures was also tested. This contains vowel F1~F3 and durations, and /s/ CoG measurements from each *just* token. Figure 5 displays a comparison between the best performing systems across segmental measurements, and this combined system.

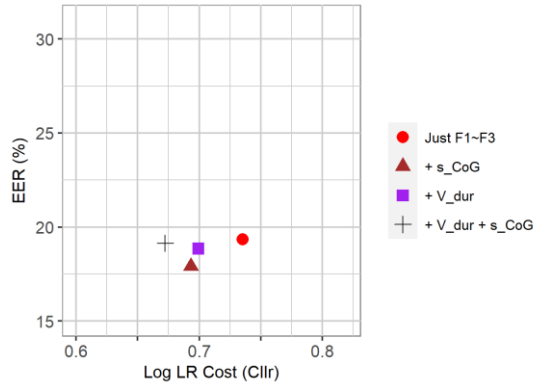

Figure 5: Mean C_{llr} plotted against mean EER comparing the best performing combinations of *just* acoustic measurements.

Overall, the combined single-parameter system performs better than F1~F3, with a mean C_{llr} of 0.67 and a mean EER of 19.1% – the lowest mean C_{llr} score across all systems tested. There are, however, very slight improvements in model performance against a system containing only F1~F3 measurements. A reduction in C_{llr} measures is seen across 19 out of the 25 replications, whereas a reduction in EER is only seen across 9 replications. The system with the lowest mean EER is the one containing F1~F3 and /s/ CoG measurements (17.9%).

4. DISCUSSION

Combining *just* single-parameter measurements provides the best system for FVC in this study. Looking at the word as a forensic feature of the voice, it is possible to combine many components to make a more robust speaker discriminant. As [2] argues, ideal speaker discriminant features of the voice should be easy to measure, but display low within-speaker and high between-speaker variation. There is a fine balance between these two requirements. On one hand, features should be as robust as possible in their discriminatory power, but often, as in the case of *just*, a higher performance is found with a complex combination of features. Vowel formant estimates are widely used in FVC analysis and casework, and that is why the baseline for much of this analysis has been F1~F3 vowel midpoint estimates. By adding /s/ CoG and vowel duration measurements, this increases model complexity but also produces the best performing system. Even when only considering vowel formants, however, *just* outperforms *um* and STRUT as a speaker discriminant in this study indicating that a word token contains more speaker-specific information than the lexical vowel it canonically relates to.

It was predicted that long-term measurements would perform better as speaker discriminants than the single-parameter (segmental) measurements. There is, in fact, a reduction in performance when including word durations and MFCCs.). It is possible that word duration is sensitive to the segmental variation of *just* in that the presence/absence of segments influenced how long or short a token would be. However, the performance of MFCCs in this study is problematic. C_{llr} scores higher than 1 are not useful for FVC, so it is surprising that MFCC measurements of *just* yielded a C_{llr} of 1.58. The problem may be to do with MVKD, which assumes that each data point is correlated even though each MFCC is not correlated. Previous research on LR-based FVC has highlighted an issue with MVKD when dealing with a high number of parameters [19]. Future research may find a better performance for word-based MFCCs in FVC systems using a less-problematic method of analysis such as Gaussian Mixture Model-Universal Background Models (GMM-UBMs).

Just shows good potential as a speaker discriminant. More work may reveal that other long-term measurements such as long-term formants and fundamental frequency contribute to its effectiveness. As it is, the potential of word tokens in FVC analysis broadens what is thought of as an individual's speech pattern beyond segmental or suprasegmental variation.

5. REFERENCES

[1] P. Rose, 'Forensic speaker identification', 2003.

[2] F. Nolan, *The phonetic bases of speaker recognition / Francis Nolan*. Cambridge: Cambridge University Press, 1983.
[3] R. Love, C. Dembry, A. Hardie, V. Brezina, and T. McEnery, 'The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations', *Int. J. Corpus Linguist.*, vol. 22, no. 3, pp. 319–344, 2017.
[4] K. Woolford, 'Just in Tyneside English', *World Englishes*, Jan. 3AD, doi: 10.1111/weng.12542.
[5] S. Tagliamonte, *Teen talk: The language of adolescents*. Cambridge University Press, 2016.
[6] J. Bybee, 'Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change', *Lang. Var. Change*, vol. 14, no. 3, pp. 261–290, 2002.
[7] F. Nolan, K. McDougall, G. de Jong, and T. Hudson, 'The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research', *Int. J. Speech Lang. Law*, vol. 16, pp. 31–57, 2009.
[8] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer'. p. [Computer program], 2022. Accessed: May 18, 2022. [Online]. Available: https://www.praat.org
[9] V. Hughes, 'Script for extracting segmental MFCCs from a series of tokens saved as individual sound files'. 2018.
[10] R Core Team, 'R: A Language and Environment for Statistical Computing.' R Foundation for Statistical Computing, Vienna. URL: https://www.r-project.org, 2020.
[11] H. Wickham *et al.*, 'Welcome to the Tidyverse', *J. Open Source Softw.*, vol. 4, no. 43, p. 1686, 2019, doi: 10.21105/joss.01686.
[12] V. Hughes, S. Wood, and P. Foulkes, 'Strength of forensic voice comparison evidence from the acoustics of filled pauses', *Int. J. Speech Lang. Law*, vol. 23, no. 1, pp. 99–132, 2016.
[13] B. Robertson and G. A. Vignaux, 'DNA evidence: Wrong answers or wrong questions?', *Genetica*, vol. 96, no. 1, pp. 145–152, Jun. 1995, doi: 10.1007/BF01441160.
[14] J. Lo, 'fvclrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison'. 2022. Accessed: Oct. 04, 2022. [Online]. Available: https://github.com/justinjhlo/fvclrr#readme
[15] B. X. Wang, V. Hughes, and P. Foulkes, 'The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison', *Speech Commun.*, vol. 138, pp. 38–49, Mar. 2022, doi: 10.1016/j.specom.2022.01.009.
[16] C. G. Aitken and D. Lucy, 'Evaluation of trace evidence in the form of multivariate data', *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 53, no. 1, pp. 109–122, 2004.
[17] G. S. Morrison, 'Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio', *Aust. J. Forensic Sci.*, vol. 45, no. 2, pp. 173–197, 2013, doi: 10.1080/00450618.2012.733025.
[18] N. Brümmer and J. Du Preez, 'Application-independent evaluation of speaker detection', *Comput. Speech Lang.*, vol. 20, no. 2–3, pp. 230–275, 2006.
[19] B. Nair, E. Alzqhoul, and B. J. Guillemin, 'Determination of Likelihood Ratios for Forensic Voice Comparison Using Principal Component Analysis', *Int. J. Speech Lang. Law*, vol. 21, no. 1, pp. 83–112, Jun. 2014, doi: 10.1558/ijsll.v21i1.83.