Front-end approaches to the issue of correlations in forensic speaker comparison

Erica Gold[1] and Vincent Hughes[2]

[1]Department of Linguistics and Modern Languages, University of Huddersfield, U.K.
[2]Department of Language and Linguistic Science, University of York, U.K.
[1]e.gold@hud.ac.uk, [2]vincent.hughes@york.ac.uk

ABSTRACT

In likelihood ratio (LR)-based forensic speaker comparison it is essential to consider correlations between parameters to accurately estimate the overall strength of the evidence. Current approaches attempt to deal with correlations after the computation of LRs (back-end processing). This paper explores alternative, front-end techniques, which consider the underlying correlation structure of the raw data. Calibrated LRs were computed for a range of parameters commonly analysed in speaker comparisons. LRs were combined using (1) an assumption of independence, (2) the mean, (3) assumptions from phonetic theory, and (4) empirical correlations in the raw data. System (1), based on an assumption of independence, produced the best validity (C_{llr} = 0.04). Predictably, overall strength of evidence was also highest for system (1), while strength of evidence was weakest using the mean (2). Both systems (3) and (4) performed well achieving C_{llr} values of ca. 0.09.

Keywords: Likelihood ratio, forensic speaker comparison, correlations, front-end processing

1. INTRODUCTION

Forensic speaker comparison (FSC) typically involves the analysis of a recording of the voice of a known suspect (e.g. police interview) and a recording of the voice of an unknown offender (e.g. bomb threat). The auditory-acoustic (AuAc) method is most commonly used for the analysis of samples in such cases [7]. The AuAc method involves a componential analysis of a wide range of segmental, supra-segmental, linguistic and non-linguistic parameters (see [5]), in which analytical listening is combined with quantification through acoustic analysis. Consistent with developments across the forensic sciences (led by DNA analysis), there is an increasing consensus that FSC evidence should be interpreted and evaluated using the likelihood ratio (LR) framework. The LR provides a gradient assessment of the strength of the evidence based on its probability under the competing propositions of prosecution and defence. Over the last 15 years there has been a considerable amount of research considering the application of the numerical LR to the evaluation of speech recordings in FSC.

However, as highlighted in [8], there remain a number of difficulties associated with the application of the fully data-driven, numerical LR approach in FSC, due to the inherent complexity of speech as a form of evidence. One such issue is how to combine LRs from individual parameters into an overall LR (OLR). Naïve Bayes [10] allows LRs from individual parameters to be combined using simple multiplication if each piece of evidence is independent of the other. Unfortunately, with naïve Bayes, there is a high risk of doubling the same evidence if correlated parameters are considered in the evaluation. This is of particular importance in FSC given that speech parameters are known to display a highly complex correlation structure due to biological, articulatory and sociolinguistic factors.

In the absence of techniques for combining LRs, early FSC LR research applied naïve Bayes irrespective of the correlations in the raw data [e.g. 15]. More recently, logistic regression fusion [4] has been used; a method developed in the field of automatic speaker recognition (ASR) for combining the results of different ASR systems. Fusion is a form of back-end processing which considers correlations in the resulting LRs rather than correlations in the raw input data. Therefore, as suggested by Rose "it is … possible … that two segments which are not correlated by virtue of their internal structure and which therefore should be naïvely combined, nevertheless have LRs which do correlate" [14]. Equally the reverse is possible, whereby correlated parameters generate non-correlated LRs.

Therefore, it is preferable, and linguistically more appropriate, to consider correlations prior to the computation of LRs. We refer to this approach as front-end processing. This paper considers alternative front-end approaches to dealing with correlations and assesses their effects on LR output. Parallel sets of LRs were computed using the same 36 Standard Southern British English (SSBE) speakers for a wide-range of phonetic parameters commonly analysed in FSC. The LRs were then combined using four methods. Firstly, OLRs were generated using a naïve Bayes assumption of independence between all parameters. This was

intended to serve as a baseline for the least conservative strength of evidence by considering all possible parameters, irrespective of existing correlations. Secondly, OLRs were calculated as the mean of the LRs for each parameter. Thirdly, phonetic theory was used to predict which parameters should be correlated in order to identify a subset of the best performing independent parameters. These were then combined using naïve Bayes. Finally, correlations between all parameters were tested empirically. This information was used to identify a subset of the best performing independent parameters, which were then combined using naïve Bayes. The systems were compared in terms of the magnitude of the resulting OLRs and their validity, evaluated using the log LR cost (C_{llr}) function [3].

2. METHOD

2.1. Database and speakers

Data were drawn from Task 1 & 2 recordings for 36 male speakers of SSBE, aged 18-25 from the DyViS [13] database. In Task 1 participants were questioned in a mock police interview. Task 2 involves the same speakers discussing the same mock crime over the telephone with an 'accomplice'. Task 2 data were extracted from the direct recording, rather than the telephone recording. Both tasks were used for the analysis of parameters based on the availability of existing data.

2.2. Parameters, features and data extraction

In order to reflect practise in real casework, a wide range of parameters were included in the analysis. Data for a number of parameters had already been extracted for the 36 target speakers as part of previous research using DyViS. The following parameters were available:

- Articulation rate (AR): mean syllables/sec [6] (Task 2)
- Fundamental frequency (f0): mean & standard deviation (SD) [6] (Task 2)
- Long-term formant distributions (LTFDs): F1~F4 [6] (Task 2)
- Hesitation markers UH ('err') & UM ('erm'): F1~F3 midpoint (+50%) [16] (Task 1)
- /aɪ/: F1~F3 +20% and +80% point of the trajectories [9] (Task 1)

Additionally, the following parameters were added:

- Word-initial /t k/: VOT (ms) & closure duration (ms) (Task 2),
- /a uː ɔː/: F1~F3 midpoint (+50%) (Task 1).

VOT and closure durations were extracted using PRAAT by identifying the onset and offset of the hold and release phases of initial /t k/ tokens, as well as the onset of periodicity in the following vowel. Similarly, vowel tokens were hand-segmented in PRAAT. Tokens were excluded from the analysis if they occurred adjacent to liquids /l r w/ or in unstressed syllables. Following [11], formant measurements were taken at +10% steps using a script set to identify between 5 and 6 formants within a 0-5000 Hz range. The +50% measurement, the temporal midpoint, from each formant was used as input.

2.3. LR computation

The 36 speakers were divided into two sets of 18 speakers to act as development and test data. The same speakers were also used as reference data. Each dataset for each speaker was also divided in half to create two sets, which acted as mock suspect and offender data, allowing for same speaker comparisons. For each parameter, cross-validated LR scores were computed using a MATLAB implementation [12] of Aitken and Lucy's [1] multivariate kernel density (MVKD) formula for the development and test data. For each comparison, the reference data consisted of 34 speakers. Scores for the development data were used to generate calibration coefficients using logistic regression [3] which were applied to the test scores to convert them to calibrated LRs (LRs). This produced parallel sets of calibrated same-speaker (SS; 18) and different-speaker (DS; 306) LLRs for each set of input data. System validity was assessed using the log LR cost function (C_{llr}) [3], which penalises the system for the magnitude, rather than proportion, of contrary-to-fact LRs.

2.4. Systems

Four front-end approaches were used to handle correlations between parameters. For each system, different combinations of parameters were used to generate the OLRs.

2.1.1. System (1): Naïve Bayes

Following the naïve Bayes approach, OLRs were generated by taking the product of the LRs for each parameter for each comparison. This approach is expected to overestimate the strength of evidence (relative to methods which consider the correlation). It was included here to serve as a baseline for comparison with the other systems (i.e. better/worse performance).

2.1.2. System (2): Mean

In System (2), OLRs were calculated as the mean of individual LRs for each parameter for each comparison. Relative to the naïve Bayes approach (System (1)), the mean was expected to produce markedly weaker strength of evidence.

2.1.3 System (3): Phonetic theory

System (3) was based on predictions from phonetic theory about the correlation structure of the parameters analysed. Firstly, temporal parameters were predicted to be dependent on AR, such that faster speech should produce shorter segmental durations. Thus, the temporal parameters related to initial /t k/ were removed. Secondly, although source and filter information are predicted to be independent of each other, evidence from [2] suggests that f0 and F1 are correlated, particularly in Lombard speech (such as that used when speaking on the telephone). Therefore, f0 was included in the analysis and F1 omitted. F1 was removed since it is also typically compromised by telephone transmission in forensic cases. Finally, all of the segmental vocalic formant data was expected to be correlated with the LTFDs, since the LTFDs already contain all of the segmental vowel data, providing information about the vowel system and the shape and size of the entire vowel space. The resulting system consisted of:

- AR, f0 (mean & SD), and LTFD (F2~F4)

These parameters were modelled as multivariate data using MVKD (i.e. including all features of each parameter) and then combined, as in System (1), by taking the product of the LRs for each comparison. These parameters were combined using naïve Bayes since they were, based on predictions from phonetic theory, expected to be independent of each other.

2.1.4. System (4): Empirical correlations in the data

System (4) was based on empirical correlations calculated from the raw data itself. Mean values by-speaker were calculated for each feature of each parameter. A Spearman correlation matrix was then generated to identify features which correlate for this population. A conservative (i.e. low) accept-reject threshold of r = 0.25 was chosen to determine the independence/dependence of pairs of parameters. Such a conservative threshold was used to capture all of the meaningful correlations in the data, even if this meant assuming some features were correlated when they were not. Hierarchical cluster analysis was then used to arrange pairs of features according to the strength of their correlation. Starting with the strongest correlations, the feature with the best validity (i.e. lowest C_{llr}) was chosen for further consideration in the system. The outcome of this was ten features with the best validity, which were also empirically shown to be uncorrelated. The resulting system consisted of:

- f0 (mean), LTFD (F3 & F4) , UH (F1), UM (F2), word-initial /t/ (VOT & closure duration), /a ɔː/ (F2), and /uː/: (F3)

As with the other systems, the LRs from these individual features were combined using simple multiplication to generate OLRs.

2.5. Evaluation

The four systems were compared in terms of the magnitude of the OLRs which had been converted to log LRs (LLRs) using a base-10 logarithm. Since the distributions of LLRs are generally skewed, the median was used as a measure of the central tendency. The validity of the four systems was compared using the C_{llr} for the OLRs.

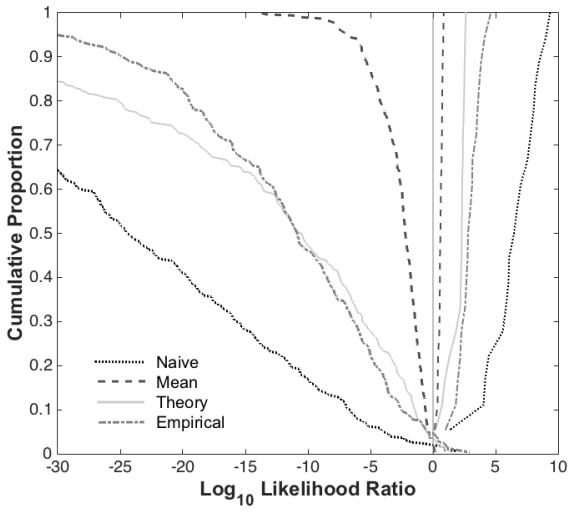
3. RESULTS

Figure 1 displays the tippett plot of calibrated overall $\log_{10}$ LRs (OLLRs) for each of the four systems tested. The largest differences were found between the naïve Bayes (1) and mean-based (2) systems, while the outputs of Systems (3) and (4) were very similar. Predictably, the highest magnitude SS and DS OLRs were produced by System (1) where all of the parameters were included and considered independent of one another. Compared with the median SS OLLR of +6.6 produced by System (1), the System (2) SS median was six orders of magnitude weaker while the medians based on theoretical (3) and empirical (4) correlations were four orders of magnitude weaker.

The differences across the systems were greater for DS pairs. The DS median for System (1) was -24.5 compared with just -2.2 for System (2), -10.7 for System (3) and -10.9 for System (4). Such differences highlight the potential for overestimating the strength of evidence when applying naïve Bayes without considering the expected, or actual, correlations in the data. Further, the low median LLRs produced by System (2) suggest that the mean provides overly conservative estimations of the strength of the evidence. Despite using different input variables, the output from the theoretical (3) and empirical (4) systems is very similar, suggesting that they both account for the correlation structure of the underlying data in similar ways.

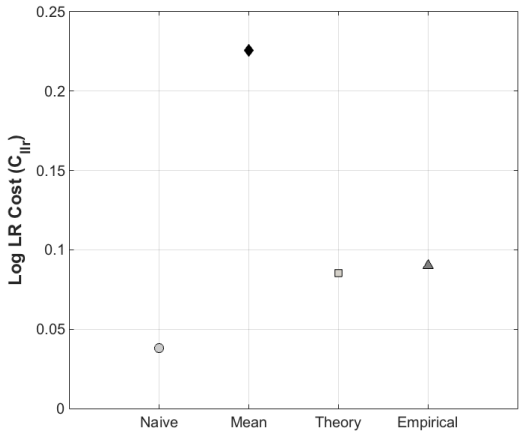
Figure 2 displays C_{llr} values for each of the four systems. Overall, the four systems outperformed any

Figure 1: Tippett plot of calibrated OLRs from the four systems using different front-end techniques to account for correlations between parameters.



single parameter in terms of validity (e.g. LTFD (F1~F4) had the best validity of any individual parameter; C_{llr} = 0.24). The system with the best validity was System (1), based on naïve Bayes (C_{llr} = 0.038). System (2), based on the mean generated the poorest validity (C_{llr} = 0.226). As in Figure 1, the output of Systems of (3) (C_{llr} = 0.085) and (4) (C_{llr} = 0.089) was very similar, although validity was marginally better when considering theory-based rather than empirical correlations.

Figure 2: C_{llr} for the four systems using front-end techniques to account for correlations.



4. DISCUSSION

As predicted, the results from the naïve Bayes system provided overestimations of the strength of evidence (see Figure 1). Statistically, System (1) offered the lowest Cllr, which is attractive in terms of system performance. However, System (1) includes portions of duplicated evidence, as pre-testing for Systems (3) and (4) showed that there are predictable correlations in the raw data which are borne out through empirical testing. From an empirical and ethical perspective, this means that System (1) may not be the most appropriate method for combining correlated evidence, despite the promising system performance.

In contrast to System (1), System (2) provides overly conservative estimations of strength of evidence, dominated by the larger number of poorer speaker discriminants (e.g. AR SS comparisons) which produce counter-factual LRs or and ones much closer to threshold.

The output of Systems (3) and (4), based on theoretical and empirical assessments of the correlation structure of the raw data, was found to be very similar in terms of both the magnitude of the OLRs and system validity. This suggests that, despite using different input parameters, these two front-end approaches account for correlations in similar ways, and to similar extents. Compared with System (1) the theory- and empirical-based approaches provide more conservative, more appropriate, assessments of strength of evidence. Further, compared with System (2), the output of Systems (3) and (4) was not overly conservative.

5. CONCLUSION

In terms of the front-end approaches currently available, we consider it preferable to account for correlations using predictions based on phonetic theory or empirical testing of the raw data since these approaches does not appear to under- or overestimate the strength of evidence in the way that naïve Bayes or the mean do. The research presented in this paper has a number of important implications for FSC. Firstly, the results highlight the complexity of the correlation structure of speech evidence and the potential effects of different front-end approaches to deal with this complexity. Secondly, the fact that the four systems outperform any single parameter in terms of C_{llr} emphasises the value of a componential approach to FSC based on all of the parameters available to the expert.

6. ACKNOWLEDGEMENTS

This research was funded in part by an International Association for Forensic Phonetics and Acoustics (IAFPA) research grant. We would like to thank the following people for their advice and support: Colin Aitken, Paul Foulkes, Peter French, and Tereza Neocleous. We are also extremely thankful to those who have aided in the collection of the data and/or lent us some of their previous DyViS results: Nathan Atkinson, Katherine Earnshaw, Natalie Fecher, Katharina Klug, and Sophie Wood.

7. REFERENCES

[1] Aitken, C.G.G., Lucy, D. (2004) Evaluation of trace evidence for discrete data. *Forensic Science International*, 230(1-3), pp. 147-155.

[2] Assmann, P.F., Nearey, T.M. (2007) Relationship between fundamental and formant frequencies in voice preference. *Journal of the Acoustical Society of America*, 122, pp. EL35-EL43.

[3] Brümmer, N., du Preez, J. (2006) Application-independent evaluation of speaker detection. Computer Speech and Language, 20(2-3), pp. 230-275.

[4] Brümmer, N. et al. (2007) Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST SRE 2006. *IEEE Transactions on Audio Speech and Language Processing*, 15, pp. 2072-2084.

[5] French, J.P., et al. (2010) The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison. International Journal of Speech, Language and the Law 17(1): 143-152.

[6] Gold, E. (2014) Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters. Unpublished; University of York. PhD.

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

[8] Gold, E., Hughes, V. (2014) Issues and opportunities: the application of the numerical likelihood ratio framework to forensic speaker comparison. *Science and Justice* 54(4): 292-299.

[9] Hughes, V., McDougall, K., Foulkes, P. (2009) Diphthong dynamics in unscripted speech. Paper presented at the *International Association of Forensic Phonetics and Acoustics Conference*, University of Cambridge, UK.

[10] Kononenko, I. (1990) Comparison of inductive and naïve Bayesian capitalised learning approaches to automatic knowledge acquisition. In B. Wielinga et al. (Eds.) *Current trends in knowledge acquisition*. IOS Press: Amsterdam, Netherlands, pp. 190-197.

[11] McDougall, K. (2004) Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law*, 11(1), pp. 103-130.

[12] Morrison, G.S. (2007) MatLab implementation of Aitken and Lucy's (2004) forensic likelihood ratio software using multivariate-kernel-density estimation. Downloaded: December 2011.

[13] 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), pp.31-57.

[14] Rose, P. (2010) Bernard's 18 – vowel inventory size and strength of forensic voice comparison evidence. *Proceedings of the 13th Australian International Conference on Speech and Technology*, Melbourne, Australia. 30-33.

[15] Rose, P., Osanai, T., Kinoshita, Y. (2003) Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. *Forensic Linguistics* 10(2): 179-202.

[16] Wood, S. (2013) Filled pauses: a discriminatory parameter for speaker comparison cases. Unpublished; University of York. MSc.