Regional variation and the definition of the relevant population in likelihood ratio-based forensic voice comparison using cepstral coefficients

Vincent Hughes, Paul Foulkes

Department of Language and Linguistic Science, University of York, UK

vh503@york.ac.uk, paul.foulkes@york.ac.uk

Abstract

This paper investigates the effects of different definitions of the relevant population with regard to regional background in LR-based forensic voice comparison using cepstral coefficients (CCs). GMM-UBM calibrated log likelihood ratios (LLRs) are computed using training and reference data in three conditions: (a) Matched, (b) Mismatched and (c) Mixed. Results suggest that there is very little validity (EER and C_{llr}) variability across conditions, with MFCCs and LPCCs equally robust to different regionally-defined populations. However, considerable variability was found in the LLRs from individual comparisons indicating that CCs encode forensically significant regional variation, which is often overlooked in ASR research.

Index Terms: regional variation, likelihood ratio, forensic voice comparison, relevant population

1. Introduction

Forensic voice comparison (FVC) typically involves the expert analysis of a recording of an unknown offender (e.g. threatening phone call) and a known suspect (e.g. police interview). Consistent with the 2009 National Research Council [1] report on strengthening forensic science and claims of a "paradigm shift" [2] across expert evidence, the likelihood ratio (LR) is increasingly accepted as the logically and legally correct framework for the evaluation of FVC evidence. The odds form of the LR is expressed as:

$$\frac{p(E \mid H_p)}{p(E \mid H_d)}, \qquad (1)$$

where p is probability, E is evidence, H_p is the prosecution proposition (same speaker) and H_d is the defence proposition (different speakers). The LR involves an assessment of the similarity between the suspect and offender samples and the typicality of the offender sample with respect to the *relevant population*.

A substantial issue for the application of the LR framework to FVC evidence is how the relevant population should be defined. In other forensic disciplines (e.g. DNA analysis), the relevant population is defined using *logical relevance*, referring to the grouping factors which affect the distribution of a variable in the population at large [3]. Applying this to FVC, Rose [4] claims that the default assumption should be that the relevant population consists of same-sex speakers of the same language as the offender (for alternatives see [5,6]). This approach has been used extensively in LR-based FVC research [7,8] and casework [9]. However, speech is an inherently complex form of forensic evidence with numerous sources of systematic between-speaker variation. Further, in FVC there is a paradox: without knowing the identity of the offender, it is not possible to know, for certain, the population of which he is a member. A small number of studies have considered how the pragmatic decisions made by analysts in defining the relevant population (e.g. which factors to control and how narrowly to control them) affect LR output for linguistic-phonetic variables [e.g. 10].

The same attention has not been given to the effects of different definitions of the relevant population on automatic speaker recognition (ASR) variables, such as CCs. The underlying structure of CCs predicts that they are not as susceptible to sources of systematic between-speaker variation as linguistic-phonetic variables [11], and therefore it may be possible to use a general database of speakers to represent the relevant population in ASR-based FVC. Consistent with this, commercially available CC-based ASR systems such as BATVOX are claimed to be "language and speech independent and thus deliver results irrespective of the language or accent used by the speaker" [12]. Similarly, a small-scale study into the effects of regional background by Moreno et al. [13] found only small equal error rate (EER) differences between matched and mismatched conditions using BATVOX, leading to the conclusion that "dialect influence is not a relevant variable for (A)SR systems … due to the fact that (A)SR uses low level acoustic characteristics not affected by differences in dialects." Despite this, evidence from Harrison and French [14] indicates that CCs are sensitive to regional variation, potentially due to known differences in vocal settings (e.g. velarised setting in Liverpool English), which are expected to affect the 'low level acoustic characteristics' analysed in CC-based ASR. However, the extent to which such regional variation affects LR output was not tested in [14].

The present paper explores the LR-based sensitivity of mel frequency (MF) and linear prediction (LP) CCs to regional background. Calibrated GMM-UBM LLRs are computed for a set of sociolinguistically homogeneous test data using different definitions of the relevant population: (a) **Matched** – where the training and reference data match the test data narrowly for regional variety, (b) **Mismatched** – involving multiple regionally homogeneous sets of training and reference data which do not match the test data, and (c) **Mixed** – where the training and reference data consist of speakers of a range of different regional varieties. In each condition, therefore, Matched, Mismatched and Mixed data was used throughout the feature-to-score conversion and score-to-LR mapping stages. Output is considered in terms of validity (EER and C_{llr}) and imprecision across relevant population conditions is assessed using 95% Credible Intervals (CIs).

2. Method

2.1. Database

Data were extracted from seven of the eight dialect regions (DRs) of the TIMIT Corpus of American English [15]. TIMIT was chosen primarily because it contains a large number of speakers (438 males; aged 21-65, mean = 31) from different regional backgrounds within a single country and language. This allows for large-scale tests with different regionally defined datasets without needing to calibrate LRs. TIMIT contains exclusively read speech in the form of 10 sentences per speaker recorded in a noise isolated sound booth. Samples were initially digitised at a sampling rate of 20 kHz and then downsampled to 16 kHz in post-production.

TIMIT is limited for the purposes of evaluating the performance of speaker recognition systems. There is a relatively small amount of data available for each speaker. Further, the samples are of a high quality, recorded in a studio in a single session, which does not reflect typical casework conditions (low quality, transmission mismatched, non-contemporaneous recordings of spontaneous speech). However, none of the available alternative databases fulfilled the essential requirements of a large number of speakers controlled for regional background within a single language. Therefore, despite the limitations of TIMIT, it does allow for the research questions in this study to be tested.

DR3 (North Midland) was chosen to act as test data (mock suspect and offender samples), since this set contained the largest number of speakers (79). 25 test speakers were firstly identified at random to function as test data. From the remaining 54 DR3 speakers, 28 were identified at random to act as a set of Matched data. For each of the other six DRs 28 speakers were extracted at random to form six Mismatched sets. Six speakers were then chosen at random from the Matched set and each of the Mismatched sets to create a 28-speaker Mixed set.

2.2. Linguistic variation

A potential issue with the use of TIMIT for investigating regional variation is the extent to which the DRs represent linguistically distinct regional varieties. There is general agreement between the DR boundaries in TIMIT and those of the major urban dialect areas of North American English identified in [16], with differences primarily in the naming conventions for each region.

Evidence of differences between the DRs from [16] also makes it possible to predict, on linguistic grounds, which Mismatched conditions should be most divergent from the training and test sets (DR3). The Western (DR7) set should display the greatest (linguistic) similarity with the Matched (DR3) set – the primary differences found in the tense long high and mid vowels and in the merger of the /ɔʊ/ and /ɔ:/ lexical sets in the West. There should also be linguistic similarity between DR3 and the DR4 (South Midland) and DR5 (Southern) sets as these three regions share laxing of long high and mid vowels. The most divergent results should be found for the Northern (DR2) set due to the Northern Cities Shift and for New York (DR6), due to /r/ vocalisation and /ɔ:/ lowering.

2.3. Feature extraction

From the ten TIMIT sentences produced by each speaker, five were assigned to the suspect condition and the remaining five were assigned to the offender condition (ca. 15s/ sample). The speech-active portion of each sample for each speaker was extracted using Morrison's [17] Sound File Cutter Upper software removing silences of greater than 100ms. MFCCs and LPCCs were extracted using HTK [18]. For each sample, a pre-emphasis filter (coefficient value 0.97) was applied to the signal. The signal was then divided into frames using a 20ms Hamming window shifted at 10ms steps, resulting in 50% overlap between adjacent frames. The power spectrum of each frame was processed by applying a mel (MFC) and linear (LPC) filter bank consisting of 26 filters across the frequency range. A discrete cosine transform was fitted to the log of the filter outputs and 12 MFCCs and 12 LPCCs extracted for LR computation.

2.4. Experiments

GMM-UBM [19] scores were initially computed for SS and DS pairs for each of the 28-speaker Matched, Mismatched and Mixed datasets, where the datasets function as both training and reference data. The reference data contained all of the available data for each speaker. GMMs of the reference data and each set of suspect data were constructed using 32 Gaussians. This number of Gaussians was determined by the small amount of suspect data available and based on [20]. Scores for the training data were used to build a logistic regression calibration model [21] for each of the experimental conditions. 25 SS and 600 DS GMM-UBM scores were then computed for the test data using the Matched, Mismatched and Mixed sets as reference data. Finally, the calibration coefficients for each condition were applied to the test scores in order to convert them to calibrated LLRs.

2.4.1. Validity

For each sets of calibrated LLRs in each of the conditions, validity was assessed using both EER and log LR cost (C_{llr}) [21]. Validity was compared across conditions for each form of CC input and the overall range of validity variability compared across the different forms of cepstral input

2.4.2. Reliability

Non-parametric 95% CIs [22] were used to estimate the imprecision in the SS and DS LLRs produced for individual comparison across the three different relevant population conditions within for both MFC and LPC input. The mean 95% CI were compared across MFCCs and LPCCs, to assess differences in sensitivity to regional variation between the different forms of input data.

3. Results

3.1. Validity

Figure 1 displays C_{llr} and EER values for each of the regional dialect-based relevant population conditions based on MFCC input. EER values are spread over a very narrow range of 0.5%, with optimum performance achieved by the Mismatched(6) (New York) system (0%) and the poorest performance achieved by the Mixed system. Relative to the performance of the Matched system (0.17%), using narrowly-defined, appropriate data, EER is marginally poorer for the Mixed and Mismatched(1; New England), (2; Northern) and (4; South Midland) systems and marginally better for the Mismatched(6) and (7) (Western) systems. The range of C_{llr} variability is also very narrow, with values spread maximally over a range of 0.03. The Matched and Mismatched(6) (New York) systems generate the best C_{llr} values (0.023) conditions. In the remaining Mismatched conditions C_{llr} is marginally poorer, although the absolute differences are very small. The poorest C_{llr} performance is recorded for the Mixed system (0.054).

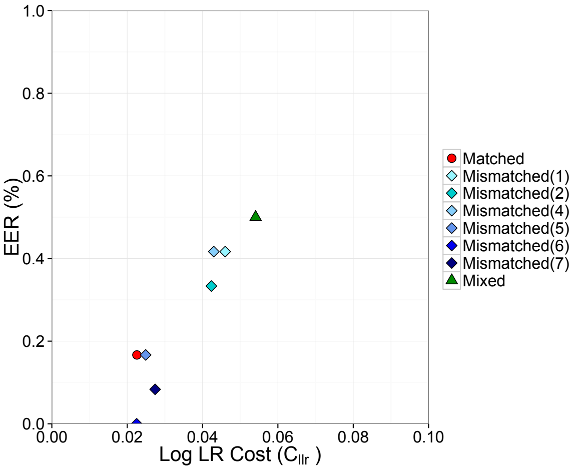

Figure 1. *Log LR Cost (C_{llr}) plotted against EER (%) for each condition based on MFCC input*

Figure 2 displays validity metrics for the eight systems based on LPCC input. EER values are spread over a range of 0.42%. Relative to the EER for the Matched system, EERs are both better and worse in the Mismatched and Mixed conditions. The best EER performance is achieved using Mismatched sets (5) (Southern) and (6) (0.083%), and the poorest performance is achieved using Mismatched sets (1) and (2) (0.5%). The Mixed system achieves the same EER as the Matched system (0.417%). A narrow range of variability is also displayed across systems in terms of C_{llr}. Values are spread over 0.039. Relative to the Matched system, C_{llr} is better for Mismatched sets (6) and (7), with (7) producing the best validity, and poorer in the other Mismatched and Mixed sets.

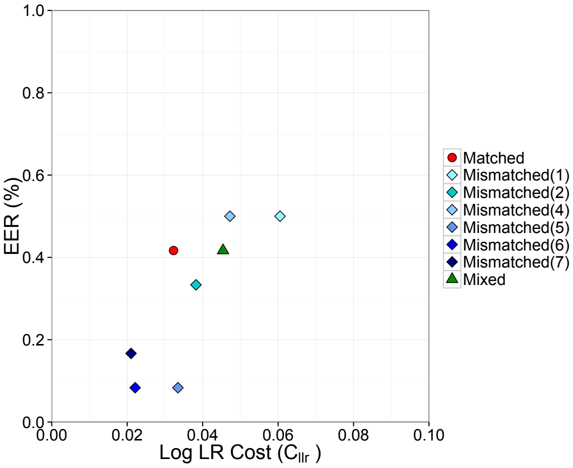

Figure 2. *Log LR Cost (C_{llr}) plotted against EER (%) for each conditions based on LPCC input*

Comparison of EER and C_{llr} values across the Matched, Mismatched and Mixed systems reveals potentially systematic patterns of variability. For both forms of input data the Mismatched sets (5), (6) and (7) are the best performing systems, consistently achieving the lowest EER and C_{llr} values. While there is variability in the ranking of the remaining sets, the Mismatched sets (1), (2) and (4) and the Mixed set consistently perform poorest. Using MFCCs, the Matched system clusters with the best performing systems, while using LPCCs it clusters with the poorest performing systems. However, the absolute differences between the systems in both Figure 1 and 2 are very small.

3.2. Reliability

Figure 3 displays the mean SS and DS LLRs (solid line) across all conditions with 95% CIs (dashed lines) for each form of cepstral input. There is almost complete overlap between the distribution of mean LLRs and 95% CIs based on MFCCs and LPCCs. This is reflected in the similarity of the mean 95% CI, which is only marginally wider using MFCCs (±1.88) than when using LPCCs (±1.80).

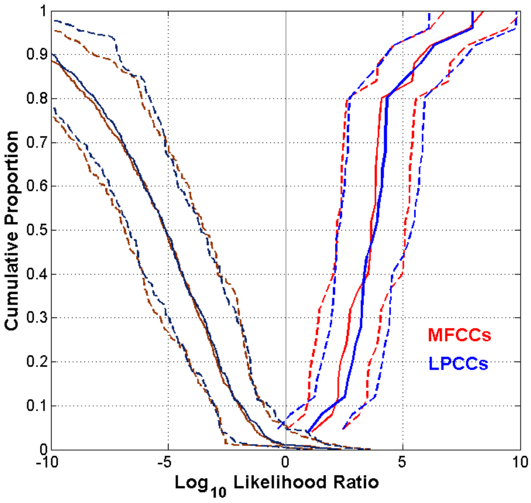

Figure 3. *Mean calibrated SS (light) and DS (dark) $\log_{10}$ LRs (solid) with 95% credible intervals (dashed) across relevant population conditions using MFCCs (red) and LPCCs (blue)*

Although the mean CIs are comparable across the different forms of input data, they do indicate a relatively large degree of imprecision in the LLRs for individual comparison pairs across the Matched, Mismatched and Mixed conditions. For example, based on MFCC input, one SS speaker comparison achieves a LLR of +7.43 using the Matched system. In the Mixed condition, the LLR for the same comparison is over two orders of magnitude stronger (+9.99) while using the Mismatched (1) data this value was stronger by three orders of magnitude. Figure 3 also suggests that the degree of imprecision across different regional-dialect system is similar for SS and DS comparisons, since the CIs for the SS and DS LLRs are roughly equal. Importantly the direction and magnitude of the variability for individual pairs is not consistent across different systems or forms of input data, indicating that different comparisons are affected by different definitions of the regionally defined relevant population in different ways.

4. Discussion

Using both MFCC and LPCC input the range of validity variability (for EER and C_{llr}) across the Matched, Mismatched and Mixed systems is found to be extremely narrow. The patterning of the performance of the Mismatched and Mixed systems is also consistent across MFCC and LPCC input with Mismatched (5), (6) and (7) sets producing the lowest EER and C_{llr} values, and the Mismatched (1), (2) and (4) and the Mixed sets generating the poorest performance. However, there is little evidence of systematic validity differences between the Mismatched/ Mixed conditions and the Matched condition, predicted by the linguistic differences between the DRs (2.2). This suggests that for LR-based MFCC and LPCC systems, validity is relatively robust to the effects of different definitions of the relevant population with regard to regional background.

However, the magnitude of the variability in individual LLRs across conditions (Figure 3) indicates that, in terms of strength of evidence, there is considerable sensitivity to the regional definition of the relevant population for MFCC and LPCC input. No systematic differences are found in the general magnitude of the imprecision for MFCC and LPCC input, or for SS and DS pairs. Rather, the use of regionally Matched, Mismatched and Mixed data affects individual comparisons in different ways and to different extents. However, due to the fact that MFCCs and LPCCs are extremely good speaker discriminants, generating strong LLRs for both SS and DS comparisons, the imprecision in the LLRs for individual speakers seemingly has very little impact on system validity. Since the data in this study was extracted from contemporaneous samples of read speech, which is expected to produce overly optimistic strength of evidence and system validity, more marked differences in validity across conditions may be revealed when using more forensically, non-contemporaneous data.

5. Conclusions

This paper has explored the effects of regional variation in the definition of the relevant population for LR-based FVC using MFCCs and LPCCs. C_{llr} and EER were found to be robust to regional variation. However, relatively large differences in the magnitude of SS and DS LLRs from individual comparisons across conditions were revealed. For both validity and reliability, the range of variation was almost exactly the same using MFCCs and LPCCs. This suggests that the choice of filterbank used to extract CCs does not markedly affect the system's sensitivity to regional variation in the data using during feature-to-score conversion or score-to-LR mapping. However, these results do offer support for the findings in [14] and challenge the claims of language and regional variety independence of CCs and of CC-based ASR systems.

6. Acknowledgements

This research was funded by the UK Economic and Social Research Council. Specific thanks go to Georgina Brown and the reviewers for *SST*. Thanks also to Anil Alexander, Esam Alzqhoul, Niko Brümmer, Philip Harrison, Geoffrey Morrison and Balu Nair for scripts which this study.

7. References

[1] National Research Council, "Strengthening forensic science in the United States: a path forward", 2009. Online: http://www.nap.edu/catalog.php?record_id=12589, accessed on 27 May 2014.

[2] Saks, M. J. and Koehler, J. J., "The coming paradigm shift in forensic identification science", Science 309:892-895, 2005.

[3] Kaye, D. H., "Logical relevance: problems with the reference population and DNA mixtures in People v. Pizarro", Law, Probability and Risk 3:211-220, 2004.

[4] Rose, P., "Technical forensic speaker identification from a Bayesian Linguist's perspective", keynote paper at Odyssey 2004, Toledo, Spain, 3-10, 2004.

[5] Morrison, G. S., Ochoa, F. and Thiruvaran, T., "Database selection for forensic voice comparison", proc. of Odyssey 2012, Singapore, 74-77, 2012.

[6] Gold, E. and Hughes, V., "Issues and opportunities: the application of the numerical likelihood ratio framework to forensic speaker comparison", Science and Justice, 2014.

[7] Morrison, G. S., "Likelihood-ratio voice comparison using parametric representations of the formant trajectories of diphthongs", Journal of the Acoustical Society of America 125(4): 2387-2397, 2009.

[8] Kinoshita, Y., Ishihara, S. and Rose, P., "Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition", IJSLL 16(1):91-111, 2009.

[9] Rose, P. "Where the science ends and the law begins: likelihood ratio-based forensic voice comparison in a $150 million telephone fraud", IJSLL 20(2):277-324, 2013.

[10] Hughes, V. "The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison", Unpublished PhD thesis, University of York, UK.

[11] Rabiner, L. and Juang, B. H. J. Fundamentals of Speech Recognition, Prentice-Hall, 1993.

[12] Agnitio, "Solution brief: criminal ID", 2013. Online: http://www.agnitio-corp.com/sites/default/files/SOL_BRIEF_Criminal_ID.pdf, accessed on 27 January 2014.

[13] Moreno, A. et al., "The influence of dialects in automatic speaker recognition", paper presented at IAFPA conference, 2006.

[14] Harrison, P. and French, J. P., "Assessing the suitability of BATVOX for UK casework: part II", paper presented at IAFPA conference, 2010.

[15] Garofolo, J. S. et al., "TIMIT acoustic-phonetic continuous speech corpus", 1993. Online: http://catalog.ldc.upenn.edu/LDC93S1, accessed on 12 November 2013.

[16] Labov, W., Ash, S. and Boberg, C., "A national map of the regional dialects of American English", 1997. Online: http://www.ling.upenn.edu/phono_atlas/NationalMap/NationalMap.html#fn0, accessed on 2 April 2014.

[17] Morrison, G. S., "Sound file cutter upper", 2010. Online: http://geoff-morrison.net/#CutUp, accessed on 24 June 2013.

[18] Young, S. et al., "The HTK Book (for HTK version 3.4)", 2006. Online: http://htk.eng.cam.ac.uk/prot-docs/htkbook.pdf.

[19] Reynolds, D. et al., "Speaker verification using adapted Gaussian mixture models", Digital Signal Processing 10:19-41, 2000.

[20] Reynolds, D., "Large population speaker identification using clean and telephone speech", Signal Processing Letter IEEE 2(3):46-48, 1995.

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

[22] Morrison, G.S., Thiruvaran, T. and Epps, J., "Estimating the precision of the likelihood ratio output of a forensic-voice-comparison system", proc. of Odyssey 2010, Brno, Czech Republic, 63-70, 2010.