Traditional Forensic Voice Comparison with Female Formants: Gaussian mixture model and multivariate likelihood ratio analyses

Phil Rose Elaine Winter

Linguistics, College of Arts & Social Sciences, The Australian National University

philip.rose@anu.edu.au wintword@live.com.au

Abstract

The first likelihood ratio-based forensic voice comparison on female voices, and the first forensic use of Gaussian mixture models on traditional features, are described. A GMM-UBM LR-based comparison is performed on the first three formants of the five long /monophthongs/ of 20 General Australian English female speakers in non-contemporaneous recordings separated by one to five weeks. Comparison with logistic-regressively fused multivariate likelihood ratios from the same data shows both systems perform well, but the latter is superior in both EER and log likelihood ratio cost.

Index Terms: Forensic voice comparison, likelihood ratio, Gaussian mixture models, multivariate likelihood ratio, logistic regression fusion, female voices, formants.

1. Introduction

Females commit crimes too. But up to now, forensic voice comparison with traditional features has been tested on male voices, e.g. [6, 12]. This paper, based on data from the second author's M. A. thesis [13] presents the results of the first likelihood ratio-based discrimination using female voices.

A second point of interest in the paper is our use of by now standard back-end processing methods in automatic forensic speaker recognition, namely Gaussian mixture modeling and logistic regression fusion.

1.1. A decade of traditional likelihood ratio-based testing

The logically and legally correct framework for forensic voice comparison, just as with any identification-of-the-source problem, involves estimating the strength of the forensic speech evidence with a likelihood ratio (LR). This is simply the ratio of the probabilities of the evidence under competing defence and prosecution hypotheses, where the evidence is the ensemble of differences between suspect and offender speech samples [3, 11]. As nicely described in [7], the likelihood-ratio based framework for forensic voice comparison emerged relatively late, at the end of the nineties, and was associated with the automatic speaker recognition community in Europe and, a little later, but polygenetically, the traditional phonetics community in Australia. The suitability of formants *qua* traditional features in LR-based FVC was first demonstrated in 2001 [4], where a naïve Bayes classifier derived from Lindley's simple [5] model was used on formants in Japanese. This model assumed, rather unrealistically, equal variances, normal distribution and independence of the variables. LRs were thus estimated separately for each formant and then combined in Naïve Bayes fashion to get an overall LR. The approach worked well enough to justify further experimentation, and the intervening decade has seen quite intensive LR-based testing with increasingly sophisticated models. These have tried to accommodate the known complexities of speech variation, in particular the non-uniform variances and non-normal distributions of variables as well as any correlation between them. The two advances in this respect were the use of kernel density modeling [11] and multivariate LRs to handle correlation between variables [1, 3, 12]. More recently, thanks mostly to the Australian Research Council Grant research of Morrison, back-end processing with methods from automatic speaker recognition have been trialed. The two main advances in this respect have been the use of logistic-regressive fusion to combine LRs from different variables [8], and the use of adaptive Gaussian mixture models [10]. This paper illustrates both of these approaches and compares their results with multivariate LR analyses. The last decade has also seen experimentation with different features, the most notable in terms of results being with parametrised formant trajectories instead of point estimates of so-called formant target values [6]. Global spectral measures (i.e cepstral coefficients) have also been shown to yield greater strength of evidence than local measures (formant centre frequencies) ; and long term statistical parameters *qua* separate pieces of evidence have also been shown to have considerable forensic potential.

Perhaps the most important result of all this activity has been to demonstrate that speech – whether automatically or traditionally processed – can be treated in exactly the same way as DNA to aid the court in determining whether the suspect did, in fact, say the incriminating speech [3].

The cornerstone of LR-based FVC is testing. Quite rightly so, given the *Daubert* requirements on admissibility of scientific evidence [2] and the new paradigm in forensic science of which LR-based FVC is now very much a part [7].

One of the ways of quantifying the strength of evidence to be expected from features used in forensic voice comparison is to use likelihood ratio-based discrimination. This approach is designed to provide an analogue to the typical scenario in forensic voice comparison and be compatible with the proper evaluation of forensic identification-of-the-source evidence, as, for example in DNA profiling [3]. Non-contemporaneous speech samples are first collected from a set of speakers, and features extracted from comparable linguistic units. A likelihood ratio is then used as a discriminant function to see to what extent the non-contemporaneous samples from the same speaker can be discriminated from those of different speakers. This is the approach used in this paper to test how well two samples from the same female speaker can be distinguished from two samples from different speakers on the basis of their vowel formant centre-frequencies.

2. Procedure

2.1. Speakers

Twenty females were recruited, all friends of the second author and all first language speakers of AE with General Australian accents. All except two, who had lived in Australia since the ages of 5 and 8 years were native-born Australians. None exhibited any observable speech defects. They were not

homogeneous with respect to age, with seven subjects in the 18-30 age group, seven in the 31-40 and six in the 41-59 group. To the extent that female accents of GA differ with respect to age, this may have aided the discrimination with pairs of samples from different speakers.

2.2. Corpus and elicitation

As appropriate in a pilot experiment like this, the data were clean. To make the study comparable with previous LR-based experiments, e.g. [12], the five long AE /monophthongs/ /i: ə: a: o: ʉ:/ were used, embedded in /h_d/ words. A set of sentences was constructed to contain the words in a stressed environment, e.g. *That's not what I heard. This seat is very hard*. Two test recordings were performed to check that speakers did indeed put the sentence stress on the target word when they read them. Participants were then recorded reading out the sentences.

Subjects were recorded on two separate occasions, separated from one to five weeks. Since speakers typically show greater variation across sessions than within (and therefore different-session within-speaker data is more difficult to discriminate), this is an important aspect of the experiment. Forensic reality also demands that suspect and offender speech samples are from different sessions. Eleven tokens of each vowel were recorded in each session. From these 22 tokens the two with the poorest extracted acoustics were discarded, leaving a total of ten replicates per session.

The samples were recorded with an Olympus DS-2200 digital voice recorder with an internal microphone. The recordings were made in low ambient noise surroundings, usually in the recording studio of the ANU School of Language Studies or at the second author's home. Due to time constraints, a couple of the second recordings had to be made at the subject's work-place (a quiet office).

2.3. Processing

The digital recordings were downloaded onto a computer and converted with *Smart Audio Converter* from .WMA into mono .WAV files sampled at 16 kHz. All remaining front-end processing was done with *Praat*. The waveform and wide-band spectrogram of the digitized speech samples were displayed and the target vowel identified. The formant centre-frequencies were then extracted with the linear prediction analysis function and superimposed on the spectrogram, so that the adequacy of the extraction could be judged visually. *Pratt*'s recommended values for adult females of five formants below a ceiling of 5.5 kHz was used as a default setting, and this appeared to pick up nicely the extra tracheal pole often found in female F-pattern, but sometimes it was necessary to adjust the formant number.

Where possible, a mean F-pattern was extracted over a maximally homogeneous section, usually about 3 - 5 centiseconds, determined by eye. This minimizes error associated with taking a measurement from just a single point in time. Where no steady state was obvious in overall F-pattern, measurement was made at a putative F2 target.

The speakers' mean formant values were calculated separately for each session, and R code was written to facilitate visual checking of the mean measurements for each vowel (an example is shown in figure 3 in the results section). After visual checking, the within-segment means for F1 F2 and F3 of each token were then transferred into a format suitable for importing into Matlab or R.

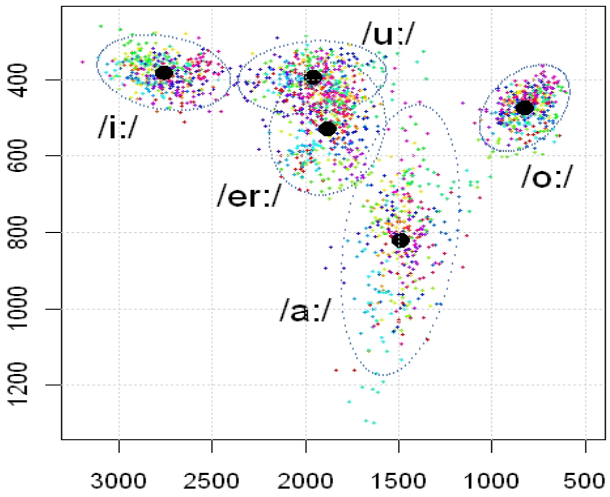

Figure 1: *F1-F2 plot of 20 speakers' five tense vowels over both recording sessions. x axis = F2 (Hz), y axis = F1 (Hz). 95% confidence ellipses are indicated.* er: = /ə/.

3. Results

Since we are using linguistic features it is appropriate to show them first. Figure 2 thus shows the mean values across both sessions for the five tense /monophthongs/ against a background of the raw data (20 vowel replicates from 20 speakers). The typical configuration presents of three central vowels /ʉ: ə: a:/ flanked by two non-low vowels /i: o:/. /ʉ:/ may have been pulled forward of central by the following alveolar consonant /d/. Compared with the configuration for 11 male speakers in [12], both the /ə:/ and the /o:/ are higher, and this perhaps contributes to the overlap between /ə:/ and /ʉ:/. Otherwise the vowels are fairly separate. These data were recently used in a real-world forensic case, to help determine which of the tense GAE /monophthongs/ a female vowel in an unclear utterance most likely belonged to.

Figure 2 shows the within-speaker variation involved by plotting the distribution of these means for the five vowels in both the F1/F2 and F2/F3 plane. The correlation structure appears different from the male data in [12]. Unlike the male speakers, who only showed correlation between F2 and F3 in /i:/ there seems to be some F2/F3 correlation in /a:/, but not in /i:/. Another difference between male and female data that may have consequences for FVC is the potential effect of a telephone bandpass. In male voices, F1 in high vowels is clearly low enough to be compromised and should not be used for comparison, but F3 is relatively immune. Figure 2 shows that very nearly all the female speakers' high vowel F1 means lie above the notional value of 350 Hz, and so F1 values may very well be able to contribute to FVC. Several of the female speakers' F3 means in /i:/, however, lie above a notional 3.5 kHz value, and this may curtail the efficacy of this particular variable in FVC with female voices.

To give a better idea of the kind of between-speaker variation in the data, and the within-speaker variation as a function of non-contemporaneity, figure 3 zooms-in on the speakers' /i:/ F1/2 means plotted separately for both sessions (this was the kind of plot used to visually check the mean measurements of all vowels). These are the data to be discriminated, and they are typical. One expects the two means from a single speaker to lie fairly close, and so they do for many speakers, e.g. speakers 1 and 17 at the bottom. Many

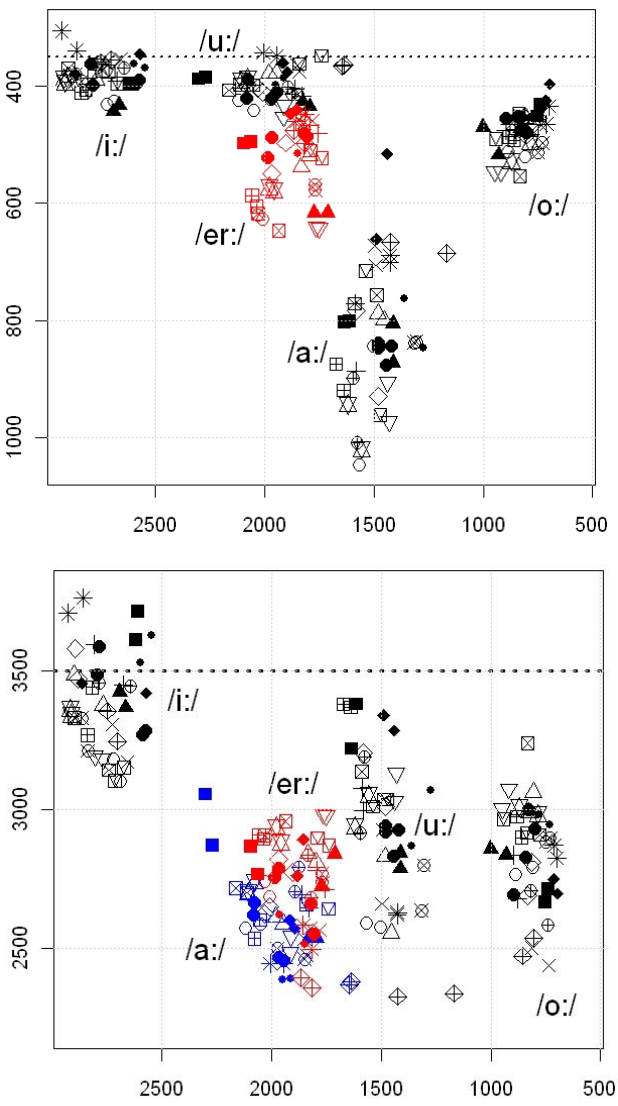

Figure 2: *Plots of F1/2 (top) and F2/3 (bottom) for 20 female speakers' five tense vowel means for both recording sessions. x axis = F2 (Hz), y axis = F1 (Hz) (top); x axis = F2 (Hz), y axis = F3 (Hz) (bottom)* er: = /ə /. *Dotted lines shows limits of notional telephone bandpass.*

different speakers' data are also fairly well separated, for example speakers 19 and 20 at the right. Nevertheless there are also instances of the same speaker's non-contemporaneous data not being similar. Some speakers, e.g. speaker 8 at the top left, are further apart, and one, (speaker 18) shows a relatively enormous difference in F2 (the data for cases like this were of course double checked). Same-speaker data like this are clearly going to be evaluated as more likely had they come from different speakers. Finally there are also data from different speakers' recordings being very similar. Note how the mean of one of speaker 18's sessions (at the right) is close to that of one of speaker 20's sessions: the between-speaker difference between these two mean values is on a par with many same-speaker differences, and will be evaluated as more likely had it come from the same speaker.

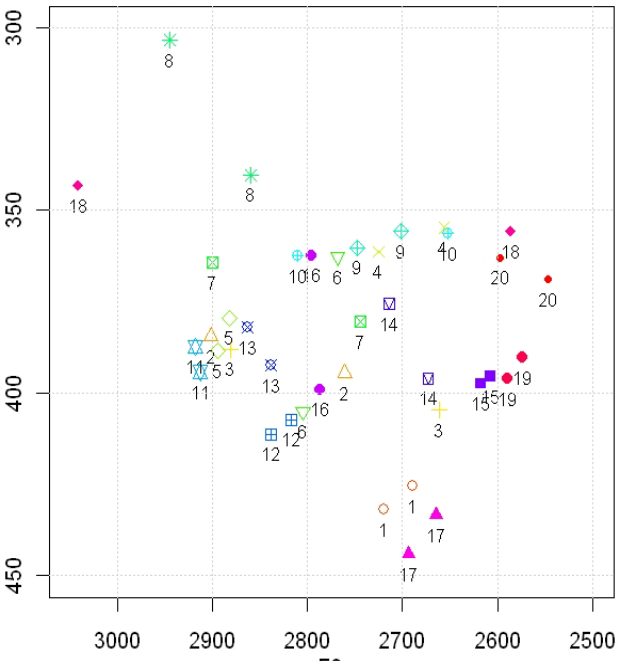

Figure 3: *F1-F2 plot of 20 speakers' /i:/ vowel means for both recording sessions. x axis = F2 (Hz), y axis = F1 (Hz).*

3.1. GMM/BM likelihood ratio

A Matlab implementation by G. Morrison of the Gaussian Mixture Model/Universal Background Model (GMM/UBM) analysis described in [10] was used. (The abbreviation *BM* for *background model* is more appropriate in this case, as the background sample could by no means be considered as 'universal'). LRs were first estimated for each vowel separately and then combined by logistic regression. The estimation of the LR for each vowel involved constructing a GMM of the three-dimensional (F1 F2 F3) data from suspect, offender and background sample. The probability of getting the offender data assuming that they have come from the suspect is then compared with the probability of getting the difference between suspect and offender data assuming the offender has come from a randomly selected member of the background sample. Because of the small number of speakers involved, cross-validation was strictly implemented at all stages of the processing. The number of mixtures and the number of iterations were varied independently (the former from 8 to 18, the latter up to 15), and an optimum combination selected on the basis of the log likelihood cost (Cllr) of the resulting system. Cllr is a hypothesis-dependent logarithmic cost function currently used for evaluating the performance of LR-based detection systems [3, 9].

The top panel of Figure 4 shows, with a Tippett plot, the results of the GMM/BM analysis on all five vowels. The curve for the different-speaker trials increases towards the left; same-speaker trials increase towards the right. It can be seen that the GMM analysis discriminates same-speaker samples from different-speaker samples fairly well, with an EER of about 5%. The Cllr is also acceptably low. The strengths of evidence attainable are also quite big, with largely symmetrical values up to about +/-$\log_{10}$ 10. This may be an effect of cross-validation on small samples, however: the LR estimate for a pair of outlying samples might be expected to increase dramatically if they are removed from the background sample, thereby overestimating the strengths of evidence attainable. One slightly worrying feature is the extent

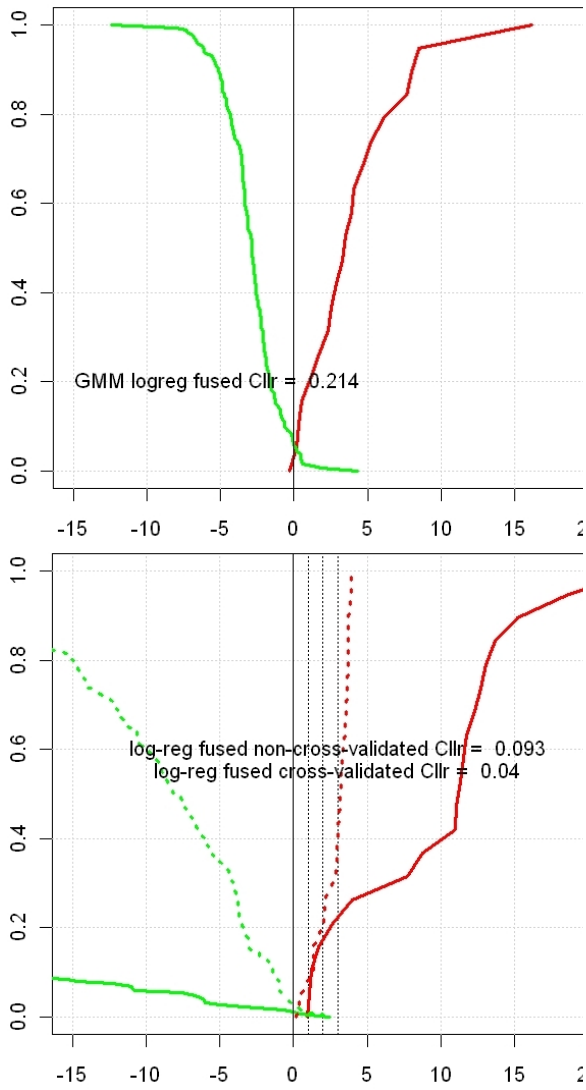

Figure 4: *Tippett plots for analysis of female vowel data. Top = GMM/BM LRs, Bottom = MVLRs. Solid lines = cross-validated LRs, dotted lines = non-cross-validated LRs. x axis = LogLR greater than …; y axis = proportion of trials.*

of the "incorrect" different-speaker comparisons, which reach up to about $\log_{10}$ 4.5. This is in fact due to just one different-speaker comparison above $\log_{10}$LR 2, between speakers 7 and 15.

3.2. Comparison with multivariate LR approaches

To assess how the GMM/BM analysis relates to previous approaches, its results were compared with an analysis using the multivariate LR approach, which is a generative LR device explicitly designed to take correlation between variables into account when estimating LRs [1]. The version of the MVLR was used which models the background sample, but not the suspect and offender samples, as a kernel density. MVLR $\log_{10}$LRs from the separate vowels were combined with logistic regression. The Tippet plot for the MVLR analysis is shown in the bottom panel of figure 4. Two sets of results are plotted, to give an idea of the effect of cross-validation. It can be seen that the logistic-regressively fused MVLRs have good discrimination, with EERs of below 1%. All same-speaker comparisons are correctly evaluated, and the largest "incorrectly" evaluated different-speaker comparison has a $\log_{10}$LR of just over 2, two orders of magnitude less than the GMM/BM results. Cllrs are very low. The previously mentioned putative effect of cross-calibration on LR estimates is confirmed: non-cross-calibration appears to rein-in the extent of achievable strength of evidence.

The reasons for the EER and Cllr superiority of the MVLR analysis presumably lie in a slightly better bias-variance trade-off for the MVLR model. It may be that there is for some reason less correlation between the MVLRs than the LRs from the GMM. Or perhaps it has to do with the MVLR use of overall, rather than specific, between- and within-speaker variance estimates. In any case none of these considerations should obscure the encouraging fact that *both* analyses resolve the female data well, and show once again the effectiveness of the LR-based approach to forensic voice comparison with speech acoustics.

4. Acknowledgements

This paper was written as part of Australian Research Council Discovery Grant No. DP0774115. Many thanks to Geoff Morrison for his code-writing expertise and contribution to FVC.

5. References

[1] Aitken, C.G.G. & Lucy, D. "Evaluation of trace evidence in the form of multivariate data", Applied Statistics 53/4, 109-122, 2004.

[2] Daubert Daubert v. Merrell Dow Pharmaceuticals, Inc. 113 S Ct 2786, 1993.

[3] Gonzalez-Rodriguez J. Rose P. Ramos, D. Torre, D. & Ortega-García, J. "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", IEEE Transactions on Audio Speech and Language Processing 15/7, pp. 2104 – 2115, 2007.

[4] Kinoshita Y: Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants. Unpublished Ph.D. Thesis, the Australian National University, 2001.

[5] Lindley DV. "A problem in forensic science", Biometrika 64/2: 207-13, 1977.

[6] Morrison, G.S "Likelihood Ratio forensic voice comparison using parametric representation of the formant trajectories of diphthongs", JASA 125, 2387-2397, 2009.

[7] Morrison, G.S. "Forensic voice comparison and the paradigm shift". Science & Justice, 49: 298–308, 2009.

[8] Pigeon, Stéphanie, Druyts, Pascal, & Verlinde, Patrick, "Applying Logistic Regression to the Fusion of the NIST '99 1-Speaker Submissions", Digital Signal Processing 10/1-3, 237-248, 2000.

[9] Ramos-Castro, D., Gonzalez-Rodriguez, J., & Ortega-Garcia, J. "Likelihood Ratio Calibration in a transparent and Testable Forensic Speaker Recognition Framework". Proc. IEEE Odyssey, 2006.

[10] Reynolds, Douglas A., Quaterieri, Thomas F., & Dunn, Robert B. "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, 10: 19-41, 2000.

[11] Rose P: "Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence". Computer Speech and Language Special Issue, 2005

[12] Rose, P. "The Effect of Correlation on Strength of Evidence Estimates in Forensic Voice Comparison: Uni- and Multivariate Likelihood Ratio-based Discrimination with Australian English Vowel Acoustics", International Journal of Biometrics 2/14: 316-329, 2010.

[13] Winter, Elaine. "Forensic speaker comparison with Australian female voices: A likelihood ratio-based discrimination using F-Pattern". Unpublished M.A. thesis, Australian National University, 2009.