FORENSIC SPEAKER DISCRIMINATION WITH AUSTRALIAN ENGLISH VOWEL ACOUSTICS

Phil Rose

The Australian National University
Philip.rose@anu.edu.au

ABSTRACT

A large-scale forensic discrimination experiment is described that investigates how well same-speaker speech samples can be discriminated from different-speaker speech samples using acoustic parameters from Australian English vowels. A multivariate likelihood ratio is used as a discriminant function on the five tense and six lax vowel phonemes of 171 male speakers. In 171 target trials and 58,140 non-target trials, comparing samples with just one token per vowel each gives EERs of between 17% and 40%, which drop to between 10% and 14% when fused. Kernel density modeling of the reference population is shown to outperform normal, and performance degrades under quasi-realistic conditions.

Keywords: Forensic speaker recognition, Bayesian likelihood ratio, vowel formants.

1. INTRODUCTION

Over about the last twenty years much attention has been given to the proper evaluation of forensic evidence, resulting in a call for a major paradigm shift in many areas of forensic identification science [11]. This has been mainly the result of the successful use of forensic DNA profiling and its way of evaluating evidence, together with some spectacular miscarriages of justice due to incorrect statistical reasoning [10]. Thus the call has now gone out for all kinds of forensic identification science to emulate DNA:

"… DNA profile evidence is now seen as setting a standard for rigorous quantification of evidential weight that forensic scientists using other evidence types should seek to emulate." [2].

If Forensic Speaker Recognition (FSR) is to emulate DNA, two things are criterial. Firstly, it must use the correct logical framework for the evaluation of evidence [6]. In FSR, all interested parties want to know the probability that two or more speech samples have come from the same speaker, given the differences, or similarities, between them. The solution is given by Bayes' Theorem, which states that the odds in favour of the hypothesis, given the evidence, is the prior odds in favour of the hypothesis times the strength of that evidence. The evidential strength is estimated by a Likelihood Ratio (LR), which is the ratio of the probabilities of the *evidence* under the competing hypotheses. LR values greater than unity support the prosecution hypothesis that the suspect said the incriminating speech; values less than unity support the defence. The magnitude of the LR is proportional to the strength of the evidence, with values close to unity meaning that the evidence is just about as likely under both hypotheses, and therefore useless. Since in FSR we do not usually know the prior odds, by Bayes' Theorem a p(Hypothesis | Evidence) statement cannot logically be given. The FSR expert must try to estimate the LR for the evidence instead [7] [8].

The second criterion is testability. According to a well-known standard for admissibility of scientific evidence in Court [4], the approach must be testable, and the error rate be known (and acceptably low). Given that the LR is predicted to be greater than unity for same-subject data, but less than one for different-subjects, it can be used as a discriminant distance around the appropriate threshold, and the evidence consisting of known same-speaker and different-speaker pairs tested to see to what extent they are correctly resolved - a relatively straightforward discrimination between same-speaker and different-speaker speech samples.

The idea of testing a theorem is not coherent since it does not possess the property of being wrong, and its truth is guaranteed. Rather, it is the discriminability of the material (here, speech), and the means of estimating its discriminability, (the LR) that is tested thereby. That is what this paper is about. We want to see how well same-speaker speech samples can be discriminated from different-speaker speech samples on the basis of their vowel acoustics - especially their formants (a common FSR parameter) - using a Likelihood Ratio.

numerator of MVLR = (1)

$$(2\pi)^{-p}\left|D_1\right|^{-1/2}\left|D_2\right|^{-1/2}\left|C\right|^{-1/2}\left(mh^p\right)^{-1}\left|D_1^{-1}+D_2^{-1}+\left(h^2C\right)^{-1}\right|^{-1/2}$$
$$\times\exp\left\{-\frac{1}{2}\left(\bar{y}_1-\bar{y}_2\right)^{\mathrm{T}}\left(D_1+D_2\right)^{-1}\left(\bar{y}_1-\bar{y}_2\right)\right\}$$
$$\times\sum_{i=1}^{m}\exp\left[-\frac{1}{2}\left(y^*-\bar{x}_i\right)^{\mathrm{T}}\left\{\left(D_1^{-1}+D_2^{-1}\right)^{-1}+\left(h^2C\right)\right\}^{-1}\left(y^*-\bar{x}_i\right)\right]$$

denominator of MVLR = (2)

$$(2\pi)^{-p}\left|C\right|^{-1}\left(mh^p\right)^{-2}\prod_{l=1}^{2}\left[\left|D_l\right|^{-1/2}\left|D_l^{-1}+\left(h^2C\right)^{-1}\right|^{-1/2}\times\sum_{i=1}^{m}\exp\left\{-\frac{1}{2}\left(\bar{y}_l-\bar{x}_i\right)^{\mathrm{T}}\left(D_l+h^2C\right)^{-1}\left(\bar{y}_l-\bar{x}_i\right)\right\}\right]$$

where U, C = within-, between-speaker variance/covariance matrices; n_1, n_2 = number of replicates per speaker
m = number of speakers in reference population; p = number of assumed correlated variables per speaker
$D_l = D_1, D_2$ = offender, suspect var/cov matrices = $n_1^{-1}U, n_2^{-1}U$
h = optimal smoothing parameter for kernel density = $\left(4/(2p+1)\right)^{1/(p+4)}m^{-1/(p+4)}$
$\bar{y}_l = \bar{y}_1, \bar{y}_2$ = offender, suspect means; y^* = $\left(D_1^{-1}+D_2^{-1}\right)^{-1}\left(D_1^{-1}\bar{y}_1+D_2^{-1}\bar{y}_2\right)$
$\bar{x}_i$ = within-speaker means of reference population.

2. DATA & VARIABLES

The Bernard corpus [3] was used for testing. Collected in the late sixties, this contains information on the F-pattern (F1-F3) and duration of the eleven monophthongal and seven diphthongal vowel phonemes of 171 male Australian speakers, assigned to one of the three earlier conventional Broad, General or Cultivated accent categories of AE. This paper examines the discriminant performance of the five tense vowel phonemes (transcribed /iː ʉː əː aː oː/), and the six lax vowels /ɪ e æ a o ʊ/. LR-based discrimination of the diphthongs /aɪ ɛɪ ɔɪ ɐʉ ɪə ɛə/ is described in [6] and [9].

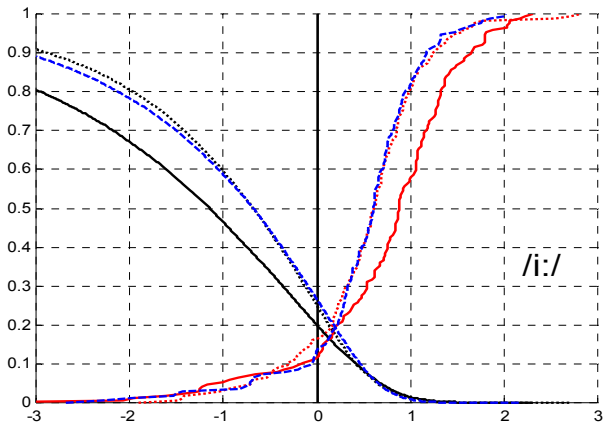
Bernard recorded his subjects saying their vowels in /h_d/ words: once with the word in isolation, once with the word in stressed sentence-final position, and once prolonged. Since many speakers were unable to produced prolonged tokens, only the isolated and sentence frame tokens are used in this paper. There are therefore only two tokens per vowel per speaker: a very tough task for discrimination. Bernard sampled F-pattern from spectrograms at onset, first target, second target if any, and offset. He also measured the duration of any onset perturbation, any target, any transition between first and second targets, and any offset perturbation.

Conditions are hardly ever optimum in real FSR case-work, and it is important to incorporate if possible some reality constraints when experimenting with forensic discrimination. Thus discrimination was carried out under two main conditions called *optimum* and *realistic*, the former making use of as many variables as possible; the latter selecting only those variables that one might expect to be able to quantify under some real case-work conditions. (*Quasi-realistic* would be a better term, since the experiment cannot, of course, simulate *every* condition encountered in the real world). *Optimum* variables were as follows. Most phonetically monophthongal vowels /əː aː oː ɪ e æ a o/ were quantified with all three formants and nuclear duration. Not surprisingly, many speakers lacked F3 for /ʊ/, and it was quantified just with F1 and F2, as well as nuclear duration. /iː/ and /ʉː/ usually have phonetically diphthongal allophones in AE and were quantified with all three formants at both targets. In addition, /iː/ was quantified with duration of its first and second targets. This was not possible with /ʉː/, where addition of duration data resulted in matrix singularity.

Since most FSR samples are from telephone speech where the F1 of high and mid vowels is compromised, the *realistic* condition omitted all F1 values at high or upper-mid targets. F3 in /o/ and

/ʊ/ is usually weak and difficult to measure, so it too was omitted (many of Bernard's speakers indeed lacked F3 for /ʊ/, although most had intact F3 for /o/). F3 was otherwise retained, since even for high front vowels it lies typically somewhat below the upper bandpass. Duration data were also discarded: they are unlikely to be of use in any but the most tightly controlled conditions. Thus, for /ɪ/ and /e/ only F2 and F3 were used; and for /iː/ and /ɯː/, F1 on the second target was omitted. F3 was omitted for /o/, and since it would only have F2 left, /ʊ/ was not used for *realistic* discrimination.

Figure 1: Tippett plot for /iː/. Vertical axis = proportion of trials; horizontal axis = $\log_{10}$LR at least ... .



3. PROCESSING

LRs were estimated with the formula derived at the *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* as a solution to the non-trivial problem of estimating the strength of evidence when predictor variables may be correlated [1]. It is vital to be able to take correlation between variables into account (and in speech many variables are), otherwise the strength of the evidence may be grossly overestimated [10].

The LR formula treats the variables for which a LR has to be estimated as multivariate data, and hence its output is called a multivariate LR (MVLR). The reference population can either be considered normal, or modeled with a Gaussian kernel density. In this paper both approaches were trialed. The formula for the kernel density version of the MVLR is reproduced from [1] at (1) (its numerator) and (2) (its denominator). The numerator quantifies the similarity between the mean values of the offender and suspect; the denominator quantifies the typicality of the difference in the reference population. The MVLR is the ratio of their values. The LR-discrimination method is intrinsic, and the same as used in [1] for testing trace evidence (elemental ratios of glass fragments). Each trial involved the comparison of one set of acoustic values from a speaker's single vowel token with either the other set of values from the same speaker or another set of values from a different speaker in the corpus. Two tokens from 171 speakers allows 171 same-speaker comparisons, or target trials, and 58,140 different-speaker, or non-target trials.

4. RESULTS

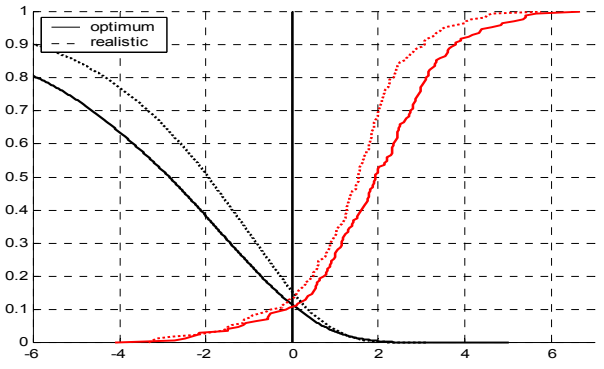
Figure 1 shows the discrimination performance of the best segment - /iː/ - in a conventional Tippett, or reliability plot. The solid lines show performance under *optimum* kernel density conditions; the dashed lines show performance under *optimum* normal, and the dotted lines show the performance with the *realistic* kernel density condition. Tippett plots are cumulative distributions of LRs from same-subject and different-subject trials (the former increase towards the right, the latter towards the left). They show for what proportion of same- or different-speaker trials one observes a LR equal to or bigger than a given abscissa LR value. This enables a clear statement of the probability of error. For example, it can be seen that ca. 20% of the different-speaker *optimum* kernel density trials (solid line) were evaluated incorrectly (as being more likely had they come from the same speaker), with a $\log_{10}$MVLR greater than the $\log_{10}$LR = 0 threshold. (This happened therefore in 11,628 out of 58,140 trials). Some 11% of the same-speaker trials (19 out of 171 trials) were also incorrectly evaluated. The equal error rate (EER) for the *optimum* kernel density condition is 17.0%. Its location close to the threshold is typical for discrimination with analytically derived LRs from many variables [9]. Figure 1 shows that greater strength of evidence, and better discrimination, is obtained with the kernel density MVLR than with the normal. This indicates that at least some of the variables in /iː/ are non-normally distributed. The decrease in strength of evidence that accompanies the *realistic* condition MVLR can also be seen, with the *realistic* and *optimum* normal curves almost congruent. This shows that F1 in the second target of /iː/, and the duration of its two targets, contain individualising information.

From table 1, which gives the EERs for all the vowels under the different experimental conditions, it can be seen that the kernel density MVLR outperforms the normal (although the overall difference is not as marked for some vowels as for the /i:/). A systematic comparison between tense and lax vowel EER is not possible, because their conditions were not all comparable. Noteworthy however is the performance of /æ/, which outperforms both comparable tense vowels /a:/ and /ə:/ in both optimum and realistic conditions.

Table 1: Equal error rates (%) for MVLR tense and lax vowel discrimination. opt = optimum, real = realistic, kden = kernel density, norm = normal.

/vowel/	kden.opt	norm.opt	kden.real
/iː/	17.0	20.0	20.0
/ʉː/	26.3	29.0	28.0
/əː/	28.0	28.1	28.2
/aː/	27.9	30.2	29.8
/oː/	28.4	29.8	40.4
/ɪ/	29.4		34.0
/e/	28.7		34.1
/æ/	**22.6**		**25.0**
/a/	28.5		31.7
/o/	26.5		35.6
/ʊ/	34.6		

Figure 2: Tippett plots for fused tense vowel MVLRs.



Finally, figure 2 shows the combined performance from the five tense vowels. This is obtained by summing the kernel density Log_{10}LRs of the individual vowels in Independence Bayes fashion. It can be seen that the EER is just over 10% for the *optimum* condition, and ca. 14% for the *realistic*. Considering they are obtained with just five vowels per sample - about 0.7 sec. of material - this is not bad, and shows the individualising potential of vocalic formant centre frequencies. Since the magnitude of the MVLR varies positively with the number of items in the sample (as it should, for less evidence must mean greater uncertainty) one would expect very much better results had the test data comprised, say, ten replicates per sample rather than just one. One factor which will have artificially aided discrimination in this experiment is the contemporaneity of the test data. The within-speaker variance for non-contemporaneous data is usually greater than within a session. The ratio of the within- to between-variance is the main determinant of the magnitude of a LR, so LRs of lesser magnitude are to be expected with non-contemporaneous comparison. Also, recall the caveats above on representativeness of the *realistic* variables: excluding F3, for example, would likewise increase the EER.

Most importantly, however, these fused EER figures represent quite substantial drops from those for the individual vowels in table 1. This reflects one of the properties of speech acoustics that makes FSR feasible: not all speakers differ from each other in the same way.

5. REFERENCES

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

[2] Baldwin, D. J. 2005. *Weight of Evidence for Forensic DNA Profiles*. Chichester: Wiley.

[3] Bernard, J.R.L. 1967. *Some measurements of some sounds of Australian English*. Ph.D. Thesis, Sydney University.

[4] Daubert 1993. U.S. Supreme Court, Daubert v. Merrell Dow Pharmaceuticals, Inc. 113 S Cr 2786.

[5] Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar M., Ortega-Garcia, J. 2006. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language* Special Issue, 20/2-3, 331–355.

[6] Gonzalez-Rodriguez, J., Rose, P., Ramos, D., Torre, D., Ortega-Garcia, J. 2006. Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. Accepted for publication in *Transactions of IEEE*.

[7] Rose, P. 2002. *Forensic Speaker Identification*, London & New York: Taylor & Francis.

[8] Rose, P. 2006a. Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence. *Computer Speech and Language* Special Issue 20/2-3, 159-191.

[9] Rose, P. 2006b. The intrinsic forensic discriminatory power of diphthongs. *Proc. 11th Australian Intl. Conf. on Speech Science and Technology*. Auckland: Australasian Speech Science and Technology Association, 64-69.

[10] Rose, P. 2006c. Accounting for Correlation in Linguistic-Acoustic Likelihood Ratio-Based Forensic Speaker Discrimination. *Proc. IEEE Odyssey Speaker & Language Recognition Workshop*, Puerto Rico: IEEE.

[11] Saks, M. J., Koehler, J. J. 2005. The coming paradigm shift in forensic identification science. *Science*, 309/5736, 892–895.