FORENSIC VOICE COMPARISON WITH MONOPHTHONGAL FORMANT TRAJECTORIES - A LIKELIHOOD RATIO-BASED DISCRIMINATION OF "SCHWA" VOWEL ACOUSTICS IN A CLOSE SOCIAL GROUP OF YOUNG AUSTRALIAN FEMALES

Phil Rose

Australian National University Emeritus Faculty

ABSTRACT

An experiment is described relating to estimation of strength of evidence in likelihood ratio-based forensic voice comparison. It is asked whether a better performance is obtained from point estimation of formant pattern targets in monophthongal vowel acoustics rather than formant trajectories. The hypothesis is tested on non-contemporaneous recordings of a custom-built challenging database of 26 young Australian female voices performing a map task. Evaluation with the log likelihood ratio cost validity metric C*llr* shows that both trajectory and target perform well, but that contrary to phonological predictions, evidence based on monophthongal F-pattern trajectory is superior to target point measurements.

Index Terms— Forensic voice comparison, likelihood ratio, F-pattern, female voices, similar-sounding speakers

1. INTRODUCTION AND AIM

In forensic speaker recognition the expert typically compares suspect and offender speech samples to help the trier-of-fact decide whether the suspect said the incriminating speech. It is now acknowledged, at least theoretically, that the expert's help should consist in furnishing the interested parties with an estimate of the strength of evidence, or likelihood ratio. This means estimating how much more likely one is to get the speech evidence – the observed differences between the known suspect and unknown offender speech samples – assuming the incriminating speech has come from the suspect (the prosecution hypothesis H_p) rather than someone else in the relevant population (the defence hypothesis H_d). This ratio of conditional probabilities of speech evidence E_{sp} under competing hypotheses $p(E_{sp} \mid H_p) / p(E_{sp} \mid H_d)$ quantifies the strength of the evidence and is the likelihood ratio (LR) [1].

In a real case, for example a $150 million dollar telephone fraud [2], the LR tells the trier-of-fact how strong the evidence is. Its other function, as in this paper, is the essential testing of the discriminability of various forensic media, e.g. DNA [3], fingerprints [4, 5], handwriting [6], SMS texts [7] and speech [8]. To illustrate this second use, figure 1, from [9], shows the cumulative distribution of LRs from 297 same-speaker and 43,956 different-speaker comparisons on non-contemporaneous landline recordings of male Japanese speakers separated by about six months. LRs from different-speaker comparisons increase towards the left; same-speaker LRs towards the right. The feature being tested is the cepstral spectrum of the five Japanese vowels parametrised by a set of LPC cepstral coefficients, and the two sets of curves – one more, one less peripheral – show the effect of

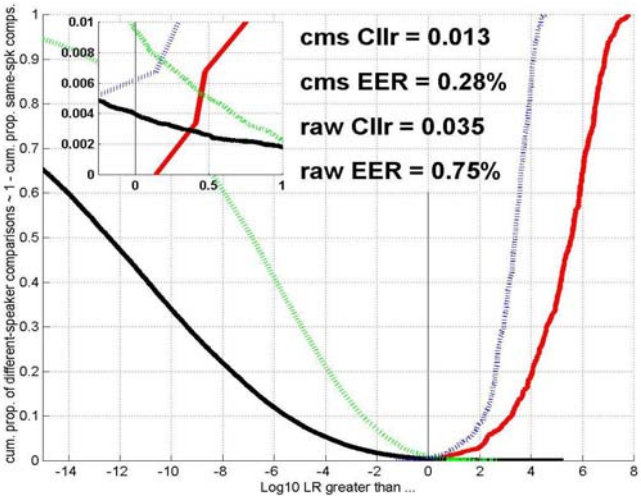

Figure 1: Tippett plot for LRs derived from comparisons using vocalic cepstral spectra. Solid lines = LRs with cepstrally-mean-subtracted-CCs; dotted lines = LRs from raw CCs. Insert shows detail around $\log_{10} = 0$. cum. prop. = cumulative proportion.

cepstral mean subtraction as a channel compensator. The excellent separation around $\log_{10}$ 0 in this so-called *Tippett* plot conveys visually that LRs based on the cepstral spectra of vowels can discriminate well between same-speaker and different speaker speech samples – read-out vowels make the task very easy, even for non-contemporaneous telephone speech – but that the performance is enhanced by cepstral mean subtraction. The performance of a system like this, equivalently its validity, is properly assessed by the information-theoretic log likelihood ratio cost C*llr* [10] which is now the metric of choice for LR-based detection systems. C*llr* values below unity indicate that the system is delivering information; features with good strength of evidence will have values well below unity. It can be seen that both the cepstrally and non-cepstrally mean subtracted C*llr*s (*cms* C*llr, raw* C*llr*) are well below unity, but the former is better. Strictly speaking the use of error rates with LRs is incorrect: by Bayes' theorem a prior probability is still required to decide whether the suspect said the incriminating speech. However, assuming flat priors for convenience, they still remain useful as indicators of discriminative power. The inset shows EERs of less than 0.1% for both features.

Emerging in the late 1990's, this logical approach to forensic speaker recognition using LRs – now often called *likelihood ratio-based forensic voice comparison* (LR-FVC) – became, after DNA, part of the new paradigm for the evaluation of forensic evidence [11]. It now seems to have survived two initial Kuhnian stages in

the emergence of a new paradigm: out-of-hand rejection and ridicule [12]. Instead, LR-FVC research has shifted emphasis from (successfully) trying to demonstrate that it can indeed emulate the so-called "DNA gold standard" [13] with both automatic and traditional approaches [14] to focusing on solving problems within the new paradigm that will enable improvements in the use of the LR in real case-work in different languages. Typical research questions address the inevitable problems with reference sample uncertainty, choice and mismatch [15, 16, 17]; and the suitability of different kinds of models and frontends [18, 9, 19, 20]. This paper is a modest example of the latter, but differs slightly in its choice of speakers.

Previous research [21, 22] has shown that, in the forensic comparison of diphthongal acoustics, where the formant pattern (F-pattern) is dynamic, better strengths of evidence are obtained if LRs are derived from F-pattern trajectories (quantified parametrically by either DCT or polynomial coefficients) rather than in traditional phonetic manner by measuring the F-pattern at two assumed diphthongal targets. However, in real-world case-work suspect and offender speech samples are also compared with respect to monophthongal vowel acoustics, and so in this paper I want to see whether a trajectory approach is also superior to a target approach for monophthongs. It should not be, given they are phonetically characterized as having a relatively unchanging transitional aspect with only a single articulatory target [23]. However, FVC requires us, of course, to be able to use the strongest features possible to compare speech samples, and this must therefore be checked.

2. SPEAKERS, ELICITATION, DATA

The speakers used in this experiment, and the way their speech was elicited, were selected to present an a priori more challenging FVC task than normal. The calculation of a LR requires a between-speaker variance estimate in order to say how likely the difference between suspect and offender is, assuming the offender is someone other than the suspect. Other things being equal, the greater the ratio of between- to within-variance, the better the forensic discriminability. The speakers were 26 (mostly) young Australian females recruited from the close friends of the research assistant involved in the experiment: from school, university, work and soccer team. Many of them knew each other. The choice of speakers homogeneous with respect to age, socio-economic background and close social group was intended to decrease the between-speaker variation and thus make the task more difficult. In addition, the elicitation sessions were all run by the research assistant, and it was expected that the enhanced interactivity between her and her mates would result in some typical in-group convergence, again contributing to a reduction in between-speaker variance. Finally, of course, female voices generally constitute more difficult investigatory objects than male, one reason being that their lesser harmonic density (from higher F0) means less acoustic information to define the spectral envelope.

A map-task was devised to elicit natural speech while still controlling the speech segments required for forensic voice comparison. Subjects were given a map, co-ordinates at its edges, of a fictitious town marked with different features – roads, buildings etc. They first had to name any features they could spot at a set of given co-ordinates by giving both the feature's name and the co-ordinate, e.g. "What's at A5?" "Hmm, A5, the Eden railway station?" Then they had to give clear instructions how to get from one map location to another, making use of the map features.

Table 1. Map task words with /ɜ/.

Map feature name	Phonemic representation	Passive articulator
Ervine theatre	ˈɜvɪn ˈɜvaɪn	zero
Erskinville lane	ˈɜskənvɪl	zero
Earthworks road	ˈɜθwɜks	zero
Burn's freeway	bɜnz	labial
BP service station	ˈsɜvəs	alveolar
Servant's boulevarde	ˈsɜvənts	alveolar
St.Mary's church	tʃɜtʃ	post-alveolar
Curtis street	ˈkɜtəs	velar

In this experiment feature names with the long mid-central monophthong phoneme /ɜ/ were used (the vowel in the word *first*). Its first three formants should lie within the nominal telephone bandwidth (0.3-3.5 kHz) and, according to Source-Filter theory, their assumed equidistant spacing will mean optimum formant amplitude. To control for prosodic context, the vowel was put in stressed word-initial position. The made-up feature names are given in table 1, where it can be seen that they were chosen to begin with onset consonants differing with respect to place of articulation: all active-articulator consonantal places (Labial, Dorsal, Coronal) are catered for, including zero. It turned out that subjects commonly used two more words with /ɜ/ in their navigation instructions: *first* /fɜst/ and *turn* /tɜn/, and these were also included, adding to the examples with Labial and alveolar onsets.

This maximum range for prevocalic consonantal place was dictated, firstly, by forensic realism: one does not usually have the luxury of controlled place of articulation in real forensic speech samples. It also has the benefit of disfavouring the trajectory hypothesis. It is well known that the F-pattern trajectory at the onset and offset phases of a vowel reflects aspects of the perivocalic segments – F2 perturbations for example reflect place of articulation. Inclusion of consonants of differing place should increase trajectory, but not target, variability (and indeed LR experiments with formant trajectories have usually controlled for the effect of prevocalic consonants). The effect should also have been reinforced by the variation in post-vocalic consonant, which was not controlled for.

Non-contemporaneous recordings are essential in FVC testing to preserve realistic within-speaker variation reflecting the details of the case [24], and 22 of the 26 subjects were recorded on two occasions separated by about one week. Elicitation was varied in the second recording to avoid learning effects.

This data acquisition process yielded ample material for testing, with each speaker having about 16 /ɜ/ replicates per session for analysis. Notable was that many participants, including the research assistant administering the map task, extensively employed creaky voice phonation, especially on low boundary tones. This phonation type has been noted for upwardly mobile American females [25] and is also commonly heard from young Australian females. Presumably it is an indexical feature signaling in-group membership and an indicator of the close social connection within the test group. Although membership of a close social network does not guarantee this, initial listening also revealed several pairs of speakers that sounded alarmingly similar:

as a naïve listener I would not have been confident in telling them apart. In view of this, it was deemed useful to test how well naïve listeners unfamiliar with the voices could generally discriminate between same-speaker and different-speaker pairs from the cohort. The phrase *A5, Eden railway station* /eɪ ˈfaɪv idən ˈreɪlweɪsteɪʃən/ was selected as stimulus for its good sampling of the speaker's peripheral acoustic vowel space (from high front in /i/ thru low central in the first target of /aɪ/ to high back rounded in /w/). Four same-speaker and six different-speaker pairs were prepared and presented to 41 volunteer listeners over the web. Forced same-speaker/different speaker decisions were required, with no priors. The test is available at [26] and was reported in [27]. It showed both same-speaker and different-speaker pairs discriminated with a wide range of accuracy: 85%, 53%, 42% and 2% (same-speaker pairs); 100%, 92%, 75%, 51%, 22% and 79% (different-speaker pairs). The overall performance (60.5% correct), although low, is frequentist-significant; but it is more appropriate to consider the results Bayesian-forensically: from the point of view of the strength of evidence associated with a same-speaker or different-speaker claim. The LR for a naïve listener's same-speaker claim (probability that a naïve listener claims that a given pair was the same-speaker, given that it was indeed the same speaker) is ca. 2.2, and the LR for a naïve-listener different-speaker claim p("it's different speakers" | different-speaker pair) is ca. 1.2. Both these strengths of evidence, approaching unity, are almost useless, and quantitatively reinforce admonitions to the legal profession about the low reliability of naïve non-familiar listeners' judgments [28].

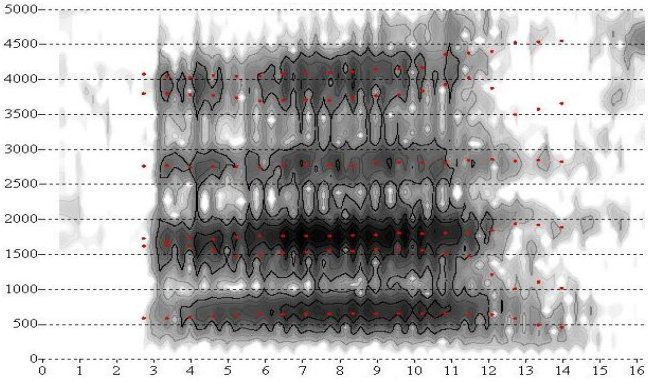

Figure 2. Spectrogram of /ɜ/ in *first* with extracted formant centre-frequencies superimposed showing extra poles below F2 and F4. X axis = duration (csec.), y axis = frequency (Hz).

3. PROCESSING

Real-world FVC case-work procedure was followed where possible. The words with stressed /ɜ/ tokens were identified aurally, the onset and offset of their /ɜ/ F-pattern determined by eye from wide-band spectrograms generated in *Praat* [29], and formants extracted. It is advisable with female voices to extract a higher number of formants than expected phonetically, as they typically have extra poles due to subglottal resonances from increased subglottal coupling [30]. This is illustrated in figure 2, which shows a wideband spectrogram, with extracted formant centre-frequencies superimposed, of the /ɜ/ vowel in a token of *first*. The presence of extra poles, which may also relate to increased subglottal coupling from the perivocalic spread-glottis voiceless fricatives [f, s], can be clearly seen in the increased bandwidth energy in the region of F4 and F2. The extracted centre-frequencies put the extra poles at ca. 3.7 kHz and 1.5 kHz. It is important to be able to discount these poles, if they are present, otherwise the LPC estimate of the F-pattern centre-frequencies will fall between the true formant and the extra pole.

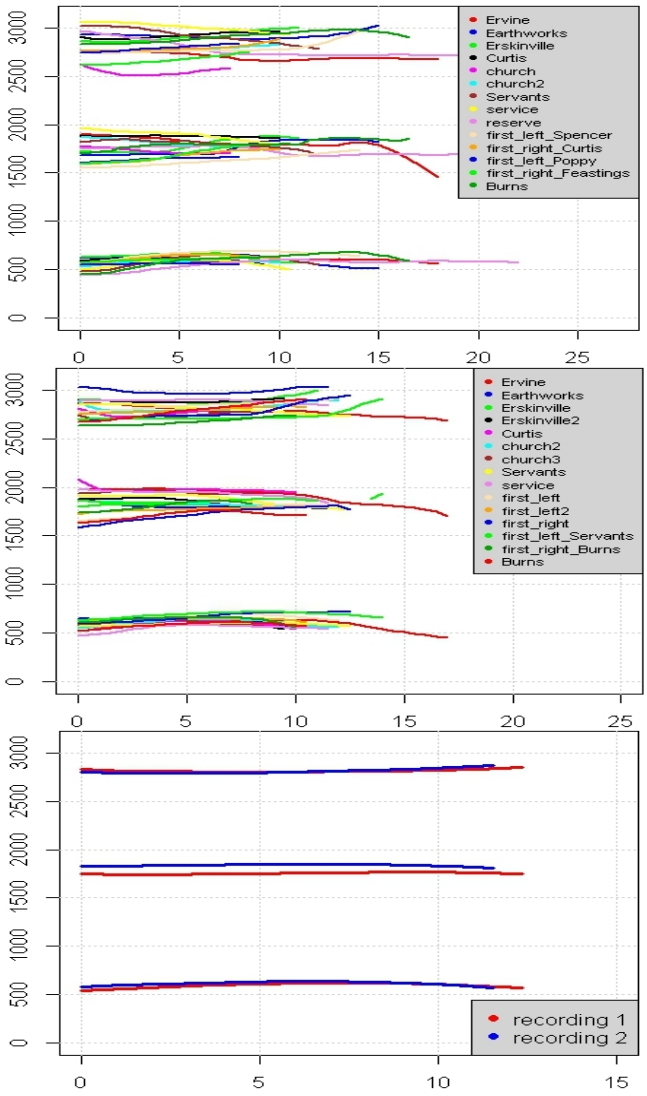

Figure 3. Stressed /ɜ/ F-pattern (Hz) as function of duration (csec.) for non-contemporaneous recordings of a single speaker. Top two panels = raw values, bottom panel = means.

Accordingly, the first six formants below 5 kHz were extracted with a specially written *Praat* script, and F1 F2 and F3 identified from them using a simple custom-written tracking code in *R* [31]. The top two panels of figure 3 show the F-pattern in both non-contemporaneous recordings of a speaker with 14 /ɜ/ replicates in the first recording and 15 in the second. Considerable variation is seen in all formants, especially F2 and F3. The bottom panel of figure 3 compares the mean /ɜ/ F-pattern of these two recordings. It can be seen that they agree rather well in their mean F1 and F3, which are almost congruent, and in their mean duration. There is a bigger difference in F2, which is about 100 Hz higher in the second recording. The good mean F-pattern resolution from what initially appears rather messy raw data is surprising, but

typical for the data. Important is that, even with the obviously simple trajectories of each formant in this bottom panel (taking their single minimum derivative point as the acoustic analog of a single articulatory target), one cannot talk of a single acoustic target at a single point in time: the putative target lies at different duration points for each formant, and for F2 at different points for the two recordings.

The raw F-pattern trajectories were modeled with cubic polynomials following [32], and their coefficients extracted for LR processing (it was confirmed that, as previously, better results are obtained when polynomials are calculated from normalised as opposed to raw duration). This resulted in (4 coefficients * 3 formants =) 12 F-pattern trajectory coefficients per vowel. Point measurements were made of putative F-pattern targets at mid- and late (75%) vowel duration, giving (3 formants * 1 target) = 3 F-pattern target features per vowel.

The test data consisted of F-pattern trajectory coefficients and F-pattern target measurements from the 22 speakers with non-contemporaneous recordings. Each speaker's first recording was compared with their second (as in figure 3) to get 22 known same-speaker LRs, and with the F-pattern of the other speakers' first recording to get 231 known different-speaker LRs.

The reference sample for estimating the between-speaker variance consisted of the 22 test data speakers together with the other four speakers who had only one recording. Since there is a considerable overlap between test and reference data, leave-one-out cross-calibration was used, whereby all data for the particular pair being tested are removed from the reference sample. The comparison was done using the multivariate kernel-density likelihood ratio (MVKD) formula developed at the *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* [33], which has been used in many previous studies as well as real-world case-work [2]. With heavily multivariate input such as the 12 trajectory coefficients, the MVKD output is often badly calibrated, and it is better to consider it not a LR but a *score* quantifying the ratio of the similarity of the difference between samples to their typicality given a suitable reference sample. It is usual then to calibrate these multivariate scores with logistic regression to convert them to true LRs [5]. This calibration was done using my *R* implementation of the *focal* toolkit [34].

4. RESULTS AND DISCUSSION

Figure 4 shows the results with a conventional *Tippett* plot (different-speaker LRs increase to the left, same-speaker LRs to the right). The fairly clear separation seen around $\log_{10}$ 0 between same- and different-speaker LRs for all three sets of features (trajectory, mid-target, late-target) indicates that the F-pattern of these young female speakers' "schwa" vowel functions quite well in distinguishing same-speaker pairs from different: all three features have C*llr*s below 0.4. I was surprised at this, given the noted auditory similarity in the cohort. It is also clear visually that the features based on the whole trajectory (solid lines) show superior validity to the target measurements, even for a monophthong, and this is confirmed in their C*llr*s: 0.29 (trajectory) vs. 0.37, 0.36 (target). EER for the trajectory LRs is 8% – 9% compared to 12% for the late target LRs, and 13% – 14% for the mid-target LRs. All these results are of course expected to worsen with transmission over telephone channels, but their relative values should remain reasonably constant.

These results show that, counter to phonetic predictions, greater strength of forensic evidence can be obtained from the trajectories of a monophthongal vowel rather than a point measurement of its target. It would appear to be the case that, even though this vowel has a single *phonetic* and *phonological* articulatory F-pattern target, speakers can still differ in how they realize it. A component of the between-speaker differences involved will also, of course, be due to differences in overall vocal tract length. Perhaps the difference between the trajectory and the target C*llr*s quantifies the amount of contribution of the trajectory component over the target component.

Given that it was intended to disadvantage a trajectory approach, the deliberately wide range of articulatory places chosen for the prevocalic consonant does not appear to have made much difference. The C*llr* from an otherwise comparable set controlled for initial consonant place of articulation would be needed to judge this.

Finally it is probably worth pointing out, in the spirit of the recent NIST *Human Assisted SR,* that, compared to the naïve listeners' values (2.2 & 1.2), the sensitivity-specificity LRs [2] of these (very limited) acoustic data are superior, especially for different-speaker judgments: $\log_{10}$LR > 0 is 9 times more likely with a same-speaker comparison; $\log_{10}$LR < 0 is 31 times more likely with a different-speaker comparison.

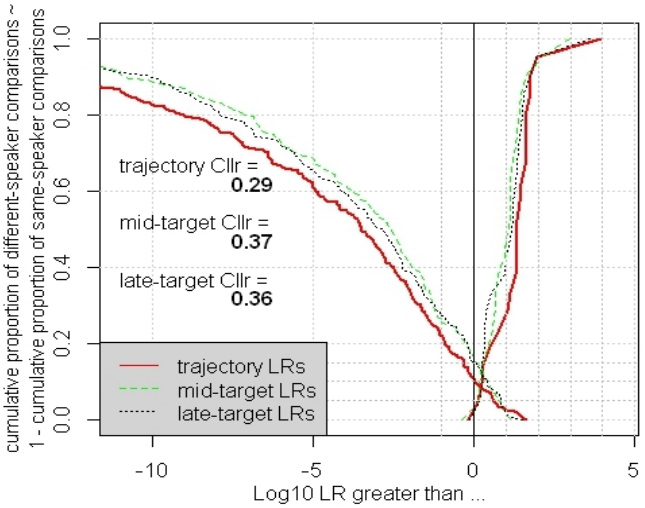

Figure 4. Tippett plot of results.

5. SUMMARY AND CONCLUSION

A simple LR-based discrimination motivated by forensic voice comparison procedure has been described on a challenging cohort of female speakers. It has shown that, counter to phonetic prediction, greater strength of forensic voice comparison evidence is obtained from the F-pattern trajectory of monophthongs, rather than from their single putative acoustic targets, even with a wide range of place of articulation of surrounding consonants. Whatever the phonological implications, this suggests that it is sensible to try where possible to base vocalic LRs on formant trajectories in real-world case-work. At least in Australian English /ɜ/, and as far as place of articulation is concerned – one would not want to extend this to differing *manners* such as /r/ or /l/, which probably still exert too great a perturbatory influence on the vocalic F-pattern.

12. REFERENCES

[1] P. Rose, "Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence", *Computer Speech and Language Special IEEE Odyssey 2004 Issue* 20/2-3, pp. 159-191, 2006.

[2] P. Rose, "Where the science ends and the law begins – likelihood ratio-based forensic voice comparison in a $150 million telephone fraud", *Int'l J. Speech Language and the Law* 20/2, pp. 277-324, 2013.

[3] I.W. Evett, J. Scrange, and R. Pinchin, "An illustration of the advantages of efficient statistical methods for RFLP analysis in forensic science", *Am. J. Human Genetics* 52, pp. 498-505, 1993.

[4] C. Neumann, I.W. Evett, and J. Skerrett, "Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm", *J. Royal Statistical Society* 175, pp. 371- 415, 2012.

[5] G.S. Morrison, "Tutorial on logistic regression calibration and fusion: converting a score to a likelihood ratio", *Australian J. Forensic Sciences*, pp. 1-25, 2012.

[6] A.B. Hepler, C.P. Saunders, L.J. Davis, and J. Buscaglia, "Score-based likelihood ratios for handwriting evidence", *Forensic Science International* 219/1-3, pp. 129-140, 2012.

[7] S. Ishihara, "A Likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams", *Int'l J. Speech Language and the Law* 21/1, pp. 23-49, 2014.

[8] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition", *Computer Speech and Language Special IEEE Odyssey 2004 Issue,* 20/2-3, pp. 331-355, 2006.

[9] P. Rose, "More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends", *Int'l J. Speech Language and the Law*, 20/1, pp. 77-116, 2013.

[10] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection", *Computer Speech and Language IEEE Odyssey 2004 Issue* 20/2-3, pp. 230-275, 2006.

[11] G.S. Morrison, "Forensic voice comparison and the paradigm shift", *Science & Justice* 49, pp. 298-308, 2009.

[12] G.S. Morrison, "Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework", *Australian J. Forensic Sciences*, 41, pp. 155-161.

[13] D.J. Balding, *Weight of Evidence for Forensic DNA Profiles,* Wiley, Chichester, 2005.

[14] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. Torre and J. Ortega-García, "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.

[15] N. Brümmer and A. Swart, "Bayesian Calibration for Forensic Evidence Reporting", *Proc. 15th Interspeech*, pp.388-392, 2014.

[16] G.S. Morrison, F. Ochoa, and T. Thiruvan, "Database selection for forensic voice comparison", *Proc. Odyssey*, pp. 62-77, 2012.

[17] S. Ishihara, "Replicate mismatch between Test/Background and Development Databases: The impact on the Performance of Likelihood Ratio-based Forensic Voice Comparison", *Proc. 15th Interspeech* pp.393-397, 2014.

[18] G.S. Morrison, "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM)", *Speech Communication* 53, pp. 242-256.

[19] P. Rose, "Forensic voice comparison with secular shibboleths - a hybrid fused GMM-multivariate likelihood ratio-based approach using alveo-palatal fricative cepstral spectra", *Proc. ICASSP* pp.5900-5903, 2011.

[20] N. Balamurali, E. Alzqhoul, and B. Guillemin, "Determination of likelihood ratios for forensic voice comparison", *Int'l J. Speech Language and the Law* 21/1, pp. 83-112, 2014.

[21] G.S. Morrison "Forensic Voice Comparison using likelihood ratios based on polynomial cures fitted to the formant trajectory of Australian English /aI/", *Int'l J. Speech Language and the Law* 15/2, pp.249-266, 2008.

[22] G.S. Morrison, "Vowel inherent spectral change in forensic voice comparison", in Morrison and Assmann (eds.), *Vowel inherent spectral change*, Springer, Heidelberg, pp. 263-283, 2013.

[23] J.M.D. Laver, *Principles of Phonetics*, University Press, Cambridge UK, 1994.

[24] E. Enzinger, "The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems", *Proc. 14th Australasian Int'l Conf. on Speech Science & Technology*, pp. 137-140, 2012.

[25] I.P. Yuasa, "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American Women", *American Speech* 85, pp. 315-337, 2010.

[26] http://staff.esteem-uc.edu.au/david/forensic-speech-task

[27] D. Vandyke, P. Rose, and M. Wagner, "The Voice Source in Forensic-Voice-Comparison: a Likelihood-Ratio based Investigation with the Challenging YAFM Database", paper presented at *International Association for Forensic Phonetics & Acoustics Conference*, 2013.

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

[29] P. Boersma and D. Weenink, "Praat: doing phonetics by computer", Version 5.1.32, http://www.praat.org/, 2014.

[30] K.N. Stevens, *Acoustic Phonetics,* MIT Press, Cambridge Mass., 1998.

[31] R Core Team. "R: A language and environment for statistical computing", http://www.R-project.org/, 2012.

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

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

[34] N. Brümmer, "Focal Toolkit", http://www.dsp.sun.ac.za/nbrummer/focal