Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English /o/ Formant Trajectories

Geoffrey Stewart Morrison [1], *Yuko Kinoshita* [2]

[1] School of Language Studies, Australian National University, Canberra, Australia
[2] School of Languages and International Studies, University of Canberra, Australia
geoff.morrison@anu.edu.au, yuko.kinoshita@canberra.edu.au

Abstract

A traditional-style phonetic-acoustic forensic-speaker-recognition analysis was conducted on Australian English /o/ recordings. Different parametric curves were fitted to the formant trajectories of the vowel tokens, and cross-validated likelihood ratios were calculated using a single-stage generative multivariate kernel density formula. The outputs of different systems were compared using C_{llr}, a metric developed for automatic speaker recognition, and the cross-validated likelihood ratios were calibrated using a procedure developed for automatic speaker recognition. Calibration ameliorated some likelihood-ratio results which had offered strong support for a contrary-to-fact hypothesis.

Index Terms: forensic speaker recognition, calibration, formant trajectories

1. Introduction

Two approaches to forensic speaker analysis have been traditional, rooted in phonetics, and automatic, rooted in engineering. This paper presents an analysis which incorporates components from both approaches. The analysis is traditional in the sense that:

- A human expert manually selected phonemes to be compared.
- Formants rather than cepstra were measured.
- Acoustic analysis was semi-automatic, with a human expert checking and, when necessary correcting, the results.
- Likelihood ratios were calculated using a generative formula which is commonly used in traditional forensic speaker recognition.

The analysis includes automatic components in that:

- The effectiveness of the system was assessed using a metric which was developed for use with discriminative automatic speaker recognition.
- The likelihood ratios were calibrated using a procedure which was developed for use with automatic speaker recognition.

2. Methodology

2.1. Data

Data consisted of laboratory-quality audio recordings of sentences of the form "Hoe, H-O-E spells hoe." The target words in the sentences all contained the phoneme /o/ (often transcribed as /oʊ/). The words were "hoe" /ho/, "Hote" /hot/, "hoed" /hod/, "bow" /bo/, "boat" /bot/, and "bode" /bod/. The sentences were read by 27 male speakers of Australian English whose ages ranged from 20 to 63 (median 39). Each speaker was recorded on two separate occasions separated by approximately two weeks. Within each session, the speaker was recorded reading each sentence twice. Written prompts were presented in random order, and the sentences containing /o/ words were mixed in with sentences containing words exemplifying a number of other vowels. Recordings were made using a Sony ECM-MS907 microphone and an Edirol R-1 recorder with the signal digitalized at 44.1 kHz.

2.2. Acoustic analysis

The first and final word of each sentence was analyzed. The beginning and end of each vowel were manually marked. The trajectories of the first three formants (F1, F2, and F3) of each word were tracked using the formant tracking procedure outlined in [1]: The number of linear-predictive-coding coefficients was fixed at nine, and formants were tracked using the algorithm described in [2]. The formants were tracked eight times using eight different cutoff values for F3 (range 2500–4000 Hz). Each of the eight formant-track sets was visually displayed overlayed on a spectrogram. The measured intensity, fundamental frequency, and formant frequencies were also used to synthesize a vowel. The researcher could listen to the original vowel and a synthesized vowel based on any desired track set. On the basis of visual and auditory inspection, the researcher selected what they judged to be the best formant-track set. The researcher also had the option of manually editing formant tracks, and of adjusting parameters for fundamental frequency measurement. Figure 1 shows the mean formant tracks averaged over all speakers and tokens.

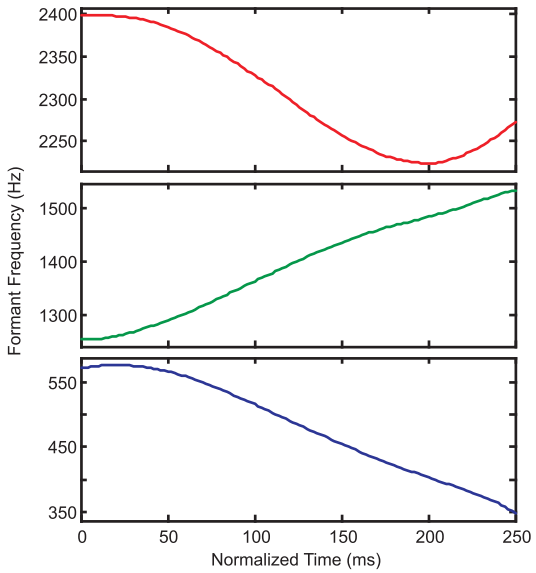

Figure 1: *Mean F1, F2, and F3 trajectories for /o/ averaged over all speakers and tokens.*

2.3. Curve fitting

Parametric curves were fitted to the three formant trajectories extracted from each vowel. Second and third order (quadratic and cubic) polynomials and discrete cosine transforms (DCT) [3,4] were fitted. The estimated coefficient values from the curve fitting were used as variables in the calculation of likelihood ratios (see §2.4). In the case of the polynomials, these included the values of the intercept plus the first through second, or first through third order terms. In the case of the DCTs, these included the values of the DC offset (zeroth coefficient) plus the first through second or first through third coefficients (the latter corresponding to the amplitudes of a half cycle, one cycle, and one-and-a-half cycles of a cosine).

Curves were fitted to trajectories scaled in hertz and in log-hertz. Also, curves were fitted to the trajectories using the original time scale (vowels varied in duration and formant measurements were available every two milliseconds) and using an equalized-duration time scale (formant values were linearly interpolated to the range 0–150 ms in 2 ms intervals).

2.4. Likelihood ratio calculation

Likelihood ratios were calculated using the multivariate kernel density formula developed by Aitken & Lucy [5,6]. This formula assesses the difference between suspect and offender samples with respect to their typicality in reference to a background distribution estimated using data from a sample taken from the appropriate population. Within-speaker variance is estimated via a normal distribution, and between-speaker variance is estimated via a kernel density model. In contrast to most automatic systems which first calculate difference scores between pairs of speech samples and use these as input to a discriminative or generative model [7], the generative Aitken & Lucy formula derives likelihood ratios via direct estimation of the probability densities of the original variables. The variables entered into the formula were the coefficient values from the parametric curve fitting (see §2.3).

Cross-validated likelihood ratios were calculated using data from all same-speaker and different-speaker pairs: Each speaker's session-one recording was compared with their own session-two recording, and with every other speaker's session-two recording. Data from all speakers except those being compared were included in the background sample. (For the calculation of the distribution of the background sample, data from both sessions were pooled within speaker.) Compared to using a model in which all data were included in the background, cross-validation provides a more realistic picture of how the system would perform on previously-unseen data, such as data from casework.

2.5. Calibration and evaluation

A typical automatic speaker recognition system consists of several stages including the calculation of difference scores between pairs of recordings, the training of a model based on the those scores, and the evaluation and calibration of the results [7]. The evaluation and calibration techniques applied to such multi-stage systems can also be applied to sets of likelihood ratios obtained from a single-stage system such as a system using the Aitken & Lucy formula [8].

The aim of calibration in forensic speaker recognition is to present the information in such a way as to best aid the finder of fact in making appropriate decisions [7]. Given two sets of values derived from two categories (such as same-speaker versus different-speaker comparisons) and a fixed decision boundary for classifying the values, calibration monotonically shifts and scales the values so as to produce the smallest possible error rate. Likelihood ratios represent the probability of obtaining the evidence under one hypothesis versus under the competing hypothesis, e. g., the probability of obtaining the observed differences between two speech samples under the hypothesis that they were produced by the same speaker versus under the hypothesis that they were produced by different speakers. The decision boundary for likelihood ratios is 1, e. g., values greater then 1 support the same-speaker hypothesis and values less than 1 support the different-speaker hypothesis (or if a logarithmic scale is used, log-likelihood-ratio values greater than 0 support the same-speaker hypothesis and values less than 0 support the different-speaker hypothesis). Minimizing error rate is equivalent to minimizing the value of a loss function. A loss function which is independent of prior probabilities and costs, and which has been adopted by the National Institute of Standards and Technology Speaker Recognition Evaluations (NIST SRE), is the log-likelihood-ratio cost (C_{llr}) [7,8,9]:

$$C_{llr}=\frac{1}{2}\left(\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}}\log_2\left(1+\frac{1}{LR_{ss_i}}\right)+\frac{1}{N_{ds}}\sum_{j=1}^{N_{ds}}\log_2\left(1+LR_{ds_j}\right)\right) \quad (1)$$

where N_{ss} and N_{ds} are the number of same-speaker and different speaker comparisons, and LR_{ss} and LR_{ds} are the likelihood ratios derived from same-speaker and different-speaker comparisons. C_{llr} is a continuous function which is small for correct likelihood ratios (same-speaker comparisons with likelihood ratios greater than 1 and different-speaker comparisons with likelihood ratios less than 1) and asymptotes towards zero as correct likelihood ratios diverge from 1, but which is large for incorrect likelihood ratios (same-speaker comparisons with likelihood ratios less than 1 and different-speaker comparisons with likelihood ratios greater than 1) and becomes exponentially large as incorrect likelihood ratios diverge from 1.

C_{llr} can be decomposed into the sum of two parts: C_{llr}^{min} is the minimum loss which can be achieved for an optimally calibrated system (a one-hundred percent correct-classification rate may be impossible because the distribution of scores from the two categories may overlap). C_{llr}^{cal} is the calibration loss, which can be reduced by shifting and scaling the scores relative to the decision boundary. C_{llr}^{min} can be used as a metric for comparing the performance of different systems.

C_{llr} and C_{llr}^{min} were calculated for the cross-validated log likelihood ratio values derived from the Aitken & Lucy formula, and using the FoCal toolkit [10] the set of log likelihood values were calibrated via a linear function optimized on C_{llr} (this results in a post-calibration C_{llr} which is somewhat greater than C_{llr}^{min} which is calculated using the non-parametric pool-adjacent-violators function). Again a cross-validation approach was adopted whereby likelihood ratios from comparisons including a given speaker were calibrated using data from all other speakers.

3. Results and Discussion

3.1. Performance of different systems

Figures 2 and 3 provide plots of C_{llr} values for the different orders of polynomial and DCT curve fitting and different combinations of time and frequency scaling. Figure 2 is based on likelihood ratios derived using F1 through F3, and Figure 3 is based only on likelihood ratios derived using F2 and F3 – in forensic-speaker-analysis casework recordings are typically made via telephone systems and the bandpass properties of telephone systems usually make F1 unusable, at least for

vowels with intrinsically low F1, such as the latter portions of /o/ (see Figure 1).

Several observations can be made on the basis of these results:

- First, in all cases C_{llr}^{cal} is relatively large indicating that substantial improvement can be achieved via calibration.
- Second, C_{llr}^{min} values for the F2 & F3 analyses were substantially larger than those for F1 through F3 analyses, indicating that F1 trajectories contain substantial information pertinent to speaker identity, and F1 trajectories should therefore be included if they are not compromised by the channel.
- Third, C_{llr}^{min} values were consistently smaller when the durations of all vowels were equalized, indicating that duration equalization improved system performance.

Beyond these observations, there appears to be relatively little difference between using polynomial and DCT curves, between using second or third order parametric curves, or between using a linear or logarithmic frequency scale. For the equalized-duration combinations, C_{llr}^{min} ranged from 0.077 to 0.092 bits for the three-formant analyses, and from 0.141 to 0.178 bits for the two-formant analyses.

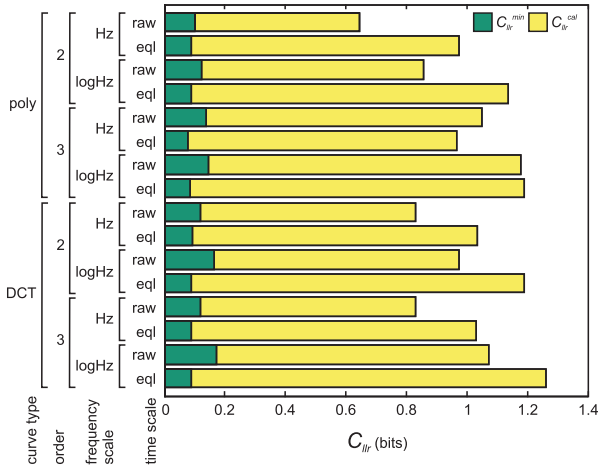

Figure 2: *C_{llr} for cross-validated likelihood ratios based on trajectories of F1, F2, and F3.*

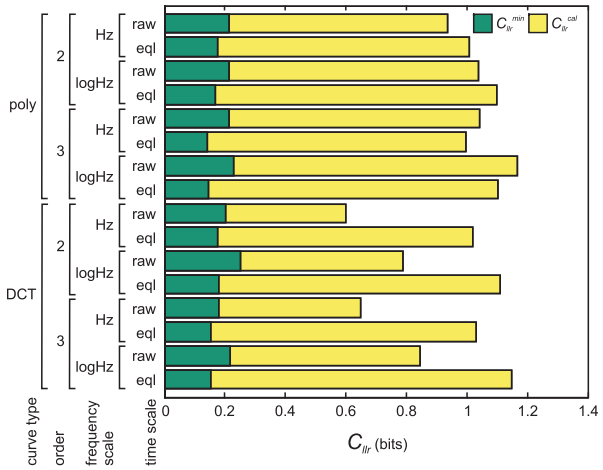

Figure 3: *C_{llr} for cross-validated likelihood ratios based on trajectories of F2 and F3.*

3.2. Calibration of best-performing system

For both the three and two-formant analyses, the best performance was achieved using third-degree polynomials fitted to linear-hertz-scaled equalized-duration formant trajectories: The three-formant system had a C_{llr}^{min} of 0.077 bits and the two-formant system had a C_{llr}^{min} of 0.141 bits. (The non-calibrated and calibrated three-formant systems had C_{llr} values of 0.966 and 0.129 bits respectively, and the non-calibrated and calibrated two-formant systems had C_{llr} values of 0.998 and 0.198 bits respectively.)

Figures 4 and 5 provide Tippett plots [8,11,12] of the non-calibrated and calibrated cross-validated likelihood ratios from the analyses based on third-degree polynomials fitted to linear-hertz-scaled equalized-duration formant trajectories.

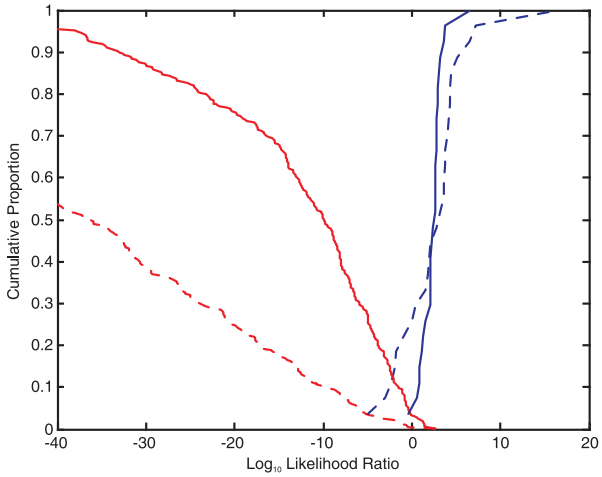

Figure 4: *Tippett plot of the non-calibrated (dashed lines) and calibrated (solid lines) cross-validated likelihood ratios from the best performing system using F1, F2, and F3 trajectories. The (red) curves rising to the left represent the proportion of different-speaker comparisons with log_{10} likelihood ratios equal to or greater than the value indicated on the x-axis. The (blue) curves rising to the right represent the proportion of same-speaker comparisons with log_{10} likelihood ratios equal to or less than the value indicated on the x-axis.*

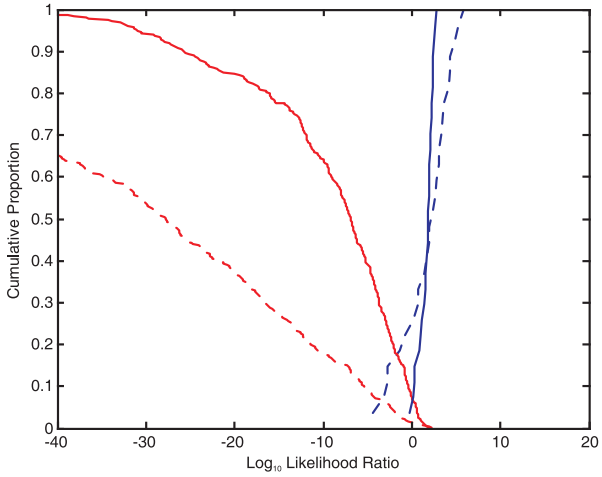

Figure 5: *Tippett plot of the non-calibrated (dashed lines) and calibrated (solid lines) cross-validated likelihood ratios from the best performing system using F2 and F3 trajectories.*

Calibration has resulted in a major reduction in misleading likelihood ratios. For example, for the three-formant system, the most misleading non-calibrated likelihood ratio from a same-speaker comparison was a likelihood ratio of 1.02×10^{-5}, i.e., 97 940 in favor of the different-speaker hypothesis. This value is of sufficient magnitude that an expert witness would probably present it in court as support for the different-speaker hypothesis, In fact it might be taken as "strong" evidence in support of the different-speaker hypothesis [13]. Since we know that in this case the two speech samples compared were both spoken by the same speaker, the presentation of such a result in court could contribute to a miscarriage of justice. In contrast, after calibration the likelihood ratio from this same-speaker comparison has shrunk to 2.21 in favor of the different-speaker hypothesis. A likelihood ratio so close to 1 would not be interpreted as meaningful support for either hypothesis.

For the three-formant system, the most misleading non-calibrated likelihood ratio from a different-speaker comparison was a likelihood ratio of 417 in favor of the same-speaker hypothesis. This was actually slightly increased to 444 in favor of the same-speaker hypothesis after calibration. Calibration has not ameliorated this contrary-to-fact likelihood ratio, but neither has it done substantial harm compared to the non-calibrated system. Using the verbal scale from [13] these likelihood ratios would be considered "moderately strong" evidence in support of the same-speaker hypothesis. This result should be considered a warning that if the procedures presented here are adopted for casework, the verbal scales of [13] should not be applied blindly. If for a comparison including a sample of unknown origin one obtains a likelihood ratio of the same magnitude as likelihood ratios which are know to be contrary to fact, then one would have to be cautious about the use of such a likelihood ratio.

Calibration has also lead to more conservative evaluations of the strength of evidence in the case of correct likelihood ratios which are far from 1. For example, for the three-formant system, the largest same-speaker likelihood ratio has shrunk from 1.47×10^{16} to 3.60×10^{6}. The number of same-speaker comparisons with likelihood ratios greater then 1000 has fallen from fourteen to five.

3.3. Poor calibration of original likelihood ratios

Theoretically Aitken & Lucy's generative formula should directly produce well calibrated likelihood ratios. The difference between the original and calibrated results is therefore somewhat surprising.

One possible reason for the poor calibration of the original cross-validated likelihood ratios could be that there is relatively little data from which to estimate the probability density functions: There were only 28 recordings of /o/ per speaker and 27 speakers, but the analysis on third-degree curves fitted to three formants required the estimation of covariance matrices for 12 parameters. Indeed bias-variance trade-offs may account for the fact that differences in performance between second-degree and third-degree parametric curves were small. This problem could potentially be remedied via the collection and analysis of a larger data set.

Another possible reason for poor initial calibration could be that the Aitken & Lucy formula does not account for all sources of variance in the speech data. The formula was originally developed for the analysis of trace evidence such as glass fragments, but the nature of speech data is more complex. In estimating the probability density of the background sample, the formula did not take into account cross-session variance. Whereas the ratios of trace elements in a pane of glass can be assumed to remain constant over time, and only sample variance and measurement error need be considered, the acoustic properties of a speaker's voice change from one occasion to another, and there may be considerable difference in a speakers voice between recording sessions. The context in which the /o/ vowels were produced is also an unaccounted-for source of variance. There was variation in the preceding and following consonant, and also in the location of the target word being at the beginning or the end of the sentence.

4. Conclusions

Applying a calibration technique developed in automatic speaker recognition to likelihood ratios derived via a traditional phonetic-acoustic analysis of Australian English /o/ formant trajectories resulted in a major improvement in the presentation of the performance of the system: Likelihood ratios purporting strong evidence in favor of a contrary-to-fact different-speaker hypothesis were shrunk to innocuous levels. However, calibration also resulted in more conservative values for large likelihood ratios which consistent-with-fact supported the same-speaker hypothesis.

5. Acknowledgments

This research was supported by the Australian Research Council grant DP0774115. Thanks to Philip Rose and three reviewers for comments on earlier versions of this paper.

6. References

[1] Nearey, T. M., Assmann, P. F., and Hillenbrand J. M., "'Evaluation of a Strategy for Automatic Formant Tracking," J. Acoustical Soc. Amer., Vol. 112, 2002, p. 2323(A).

[2] Markel, J. D. and Gray, A. H., Linear Prediction of Speech, Springer-Verlag, Berlin, 1976.

[3] Zahorian, S. and Jagharghi, A., "Speaker Normalization of Static and Dynamic Vowel Spectral Features. J. Acoustical Soc. Amer., Vol. 90, 1991, pp. 67–75.

[4] Zahorian, S. and Jagharghi, A., "Spectral-Shape Features Versus Formants as Acoustic Correlates for Vowels', J. Acoustical Soc. Amer., Vol. 94, 1993, pp. 1966–1982.

[5] Aitken, C. G. G. and Lucy, D. "Evaluation of Trace Evidence in the Form of Multivariate Data," App. Stat., Vol. 54, 2004, pp. 109–122.

[6] Morrison, G. S., Matlab Implementation of Aitken & Lucy's (2004) Forensic Likelihood-Ratio Software Using Multivariate-Kernel-Density Estimation [software], 2007. Available: http://geoff-morrison.net

[7] Ramos-Castro, D., Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems, PhD diss., Universidad Autónoma de Madrid, 2007.

[8] González-Rodríguez, J., Rose, P., Ramos, D., Toledano, D. T., and Ortega-García, J. "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition," IEEE Trans. Audio, Speech, and Lang. Proc., 15(7):2104–2115, 2007.

[9] Brümmer, N. and du Preez, J., "Application Independent Evaluation of Speaker Detection," Comput. Speech Lang., Vol. 20, 2006, pp. 230–275.

[10] Brümmer, N., FoCal Toolkit [software], 2005. Available: http://www.dsp.sun.ac.za/~nbrummer/focal/

[11] González-Rodríguez, J., Drygajlo, A., Ramos, D., García-Gomar, M., and Ortega-García, J., "Robust Estimation, Interpretation and Assessment of Likelihood Ratios in Forensic Speaker Recognition," Comput. Speech Lang., Vol. 20, 2006, pp. 331–355.

[12] Rose, P., "Accounting for Correlation in Linguistic-Acoustic Likelihood Ratio-Based Forensic Speaker Discrimination," Proc. IEEE Odyssey 2006 Speaker and Language Recognition Workshop, San Juan, Puerto Rico, June 2006.

[13] Champod, C. and Evett, I. W., "Commentary on A. P. A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification' Forensic Linguistics, 6(2):228–41," Forensic Ling., Vol. 7, 2000, pp. 238–243.