Looking into the real world: Likelihood ratio variability under forensically realistic conditions

Yuko Kinoshita

School of Culture, History and Language, the Australian National University, Australia

Yuko.Kinoshita@anu.edu.au

Abstract

This study set out to investigate how the speech of a single speaker can vary depending on their interlocutor and apparent emotional status and, consequently, how this affects likelihood ratio (LR)-based forensic voice comparison, using authentic data from past forensic casework. The results revealed that these factors do have a significant impact on the LR calculations, and voice comparisons between the testing data with mismatched conditions produce much less reliable results than those recorded under matching conditions.

Index Terms: forensic voice comparison, emotion, interlocutor, within-speaker variability, likelihood ratio

1. Introduction

Forensic voice comparison is a complex and challenging task. The research in this field has made significant progress over the last decade. An LR-based framework has been implemented, both in phonetics-based approaches and in engineering ones (e.g. [1-3]). Various features have been tested (e.g. [4-7]), LR calculation methods have been improved (e.g. [8-10]), and different approaches to evaluate the quality of LR and to post process have been put forward (e.g. [11-14]). For a comprehensive overview of these and other developments, see [15]). When it comes to casework situations, however, forensic scientists still face many problems, which are yet to be fully studied.

One of the difficulties with forensic voice comparison is the lack of control over testing data. In casework situations, scientists often have to compare two (or more) voice recordings made under very different circumstances. Recordings of unknown speakers (usually from a crime scene) are likely to be recorded in a noisy environment, with the speaker emotionally aroused—perhaps even shouting—and talking in a casual style to someone they know. Recordings of known speakers are quite often made at interviews between police officers and an apprehended suspect. The mood of these speakers tend to be nervous or depressed or at least unhappy; and they often use a different speaking style from when they talk normally to people they know. We expect these different conditions to influence speech acoustics, but we still don't know to what extent this affects the LR-based forensic voice comparison. "Are those recordings really comparable?" is almost the first question forensic scientists should ask themselves when faced by casework data, but it is rare that we can answer with full confidence. Should we just refuse to do any analysis, if we are not sure about the comparability; or is there still something that we can usefully do? Where is the useful limit for the recording conditions to be regarded as comparable? These questions must be answered, or at least be considered carefully.

Being motivated by the author's experiences in voice comparison casework as a forensic scientist, this study revisits a set of past casework data with permission from the relevant bodies. The dataset consists of 15 telephone calls made by a known speaker, talking to various interlocutors in varying emotional states. Although what we can conclude from studying a single speaker is limited, this study aims to provide a useful starting point for tackling the questions presented above.

2. Procedure

2.1. Data

2.1.1. Recordings

The testing data for this study consists of 15 phone calls made by an adult male of known identity. These calls were made from an Australian remand centre (detention centre for unconvicted suspects) over three days. They are completely spontaneous, and the physical circumstances of the speaker were consistent across these 15 phone calls. Other conditions such as the speaker's speaking style, emotional state and interlocutor varied widely.

The duration of the phone calls ranged from 53 to 720 seconds, altogether over 100 minutes. A substantial proportion was the target speaker's speech. The target speaker was bilingual in Australian English and Arabic, and switched freely between the two languages. This pilot study limits the analysis to the sections spoken in English. His English did not have discernable characteristics typical of a non-native speaker of English.

The recordings were labelled for interlocutors, and emotional states, as well as for word boundaries and segments.

Interlocutors influence our speaking style (e.g. [16]), and hence speech acoustics. The target speaker talked with a range of interlocutors including family members and other male and female speakers who seem to have a close personal connection with him, such as friends, cousins and co-workers.

Emotional states were labelled based on the author's auditory impressions of the speaker's perceived emotional states. Using auditory impressions from a panel of listeners is a scientifically rigorous approach, but the agreements for the data use did not allow this. The author's auditory impressions were supplemented by additional clues from linguistic and non-linguistic cues, such as laughter, yelling, and appearance of strongly abusive words. The labels used in classifying emotional states and interlocutors are summarized in Table 1.

This categorization of the emotional states may appear too fine grained. They do not reflect author's confidence in such classification, but they are to avoid forced categorization. This way, groups whose classification appears unreliable can be excluded from the analysis at a later stage.

Table 1. Summary of the variables

Variables	Types
Emotional states	happy, friendly, neutral/friendly, neutral, neutral/irritated, irritated, irritated/angry (in the order of positiveness of the emotion)
Interlocutors	mother, father, wife, child, male, female

2.1.2. *Feature extraction*

The acoustic quality of the recordings was sufficient for formant extraction for most part, but telephone bandwidth effects were observed. The power spectrum revealed that spectral energy was suppressed in the region below 300Hz and above 3400Hz. Therefore F1 of some high vowels would be affected.

F1–F3 of 13 phonemes of Australian monophthongs (/æ/, /ɐ/, /ɐ:/, /e/, /e:/, /ə/, /ɜ:/, /i:/, /ɪ/, /ɔ/, /o:/, /ʉ:/, /ʊ/) were extracted. Praat was used for tagging of the target vowels and extracting formants. F1 to F3 were sampled at the mid point of the vowel duration semi-automatically, and the values were checked against expected values for those formants, referring to their distributions in preceding studies [17-19]. Where the extracted values significantly deviated from the expected range, they were re-measured and manually corrected as required. The vowel space of this speaker is presented in Figure 1. Each vowel indicates the mean, and the ellipses mark the 95% range of the distribution. Each vowel spreads widely compared to data such as Bernard's [17] in which the utterances were much more controlled. F2 of /ʉ:/ varied particularly widely and it had a bi-modal distribution.

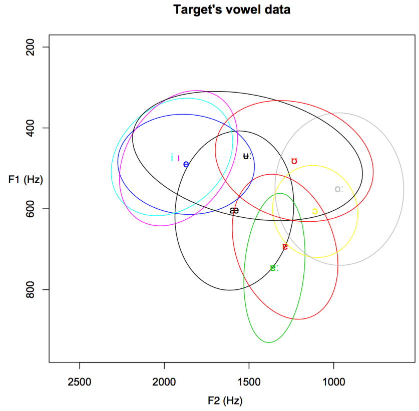

Figure 1 Target speakers vowel plot

2.2. Statistical analysis

The dataset was firstly analysed by multivariate analysis of variance to examine whether the two factors, emotional state and interlocutors, have statistically significant effects. The formant/vowel combinations found to be affected in question were further analysed.

Then multivariate kernel density likelihood ratios (MVKD) [9] were calculated for the testing samples with various mismatched conditions, and the effects of the different interlocutors and emotional states were examined. Due to the lack of data comparable to that in this study, no calibration was performed.

For the background population, this study used Bernard's formant data [17]. This dataset was extracted from recordings made under strictly controlled studio conditions in the 1960s, clearly not ideal for evaluating the strength of evidence for the testing data used here. In general, mismatch of conditions between a background population and test data makes the testing samples seem more atypical, and is likely to produce more extreme values as the significance of their similarity (or dissimilarity) will be overrated. For a same-speaker comparison, we expect a $\log_{10}$LR (LLR) greater than 0. Since this study aims to examine under which conditions the LRs deviate from this expectation rather than examining the strength of evidence itself, Bernard's formant dataset was used for its size and availability.

For the calculation of LRs, the dataset was divided into groups based on two factors: interlocutors and emotional states. For the examination of the effect of emotional states, five subgroups were selected: *neutral_friendly*, *neutral*, *neutral_irritated*, *irritated*, and *irritated_angry*, based on the availability of sufficient vowel data (183, 262, 257, 190, and 331 respectively). They were compared against each other, resulting in 10 different emotional state combinations.

For interlocutor effects, we selected three speakers: *wife*, *male* friends (which includes multiple individuals), and his young *child*. Again, these were selected because they had sufficient numbers of vowels (757, 454, and 31 respectively).

To have baseline LRs, the datasets with matching conditions were also compared. Interlocutor *wife* was selected for this purpose, as it had the largest amount of vowel data. This dataset consisted of four different emotional states: *neutral, neutral_irritated, irritated,* and *irritated_angry*; from neutral to more negative and strong. These emotion groups were further divided into two, and these two sets were compared to observe the LRs where the testing data's conditions were strictly matched.

In all cases, LRs were calculated for each vowel separately, using measurements for F1, F2 and F3; except for /i:/ and /ɪ/, where F1 was excluded from analysis as telephone bandwidth would have affected it.

3. Results

3.1. ANOVA results

Of 13 vowels, /ə/ was excluded from the analyses as there were only 3 tokens. Table 2 summarises the results of ANOVA, revealing that the factors in question affected 11 out of 12 vowels. No statistically significant effect on /e:/ was found. Thus the subsequent analysis will focus on 11 vowels excluding /ə/ and /e:/.

Table 2. Summary of ANOVA

Level of confidence	Variables	Vowels
95%	emotion, interlocutors both	ɐF2, ɐ:F1, eF2, ɪF2, ɔF1, ɔF3 æF3, eF3, iF2, ɔF1, o:F3 ɐF3, ʊF2
99%	emotion	æF1, ɜ:F2, ɜ:F3, iF1, iF2, ʉ:F2
>99.9%	emotion interlocutors	ɐ F1, ɔF2, ʉ:F1, ʊF1 ʊF2

Tukey's HSD post hoc test suggests that, in general, comparisons involving combinations with negative emotional states yield greater acoustic differences than neutral, positive or even mixed ones. For the combinations of interlocutors, *wife* and *male* produced more statistically significant difference. However, the amount of data available for these two interlocutors may be the reason for this, rather than their characteristics, as they were substantially larger than *child* dataset.

3.2. LR testing

3.2.1. Initial observations

Figure 2 below presents the 90% distributions range of the obtained $\log_{10}$LRs (LLRs). The results for all vowels were combined. The black line shows emotional state mismatch; the red line shows interlocutor mismatch; and the blue lines show matching conditions. The x-axis has been truncated for ease of observation.

Perhaps due to the significant difference in the recording conditions between the test data and Bernard’s data, some alarmingly large LLRs were observed. They supported the hypothesis consistent with the identity of the speaker (the same-speaker hypothesis), but their size suggests they are unreliable. Where a suitable training dataset for calibration is available, such spurious values may be less of a problem. However, the fact that the comparisons with matching conditions produced no such values may be noteworthy. Mismatch of conditions in both emotional states and interlocutor appears to introduce unpredictability in the LR calculation.

Also, although the three distributions substantially overlap with one another, the locations of their peaks suggest that comparisons under matching conditions tend to produce on balance slightly better evidence than the ones with condition mismatch.

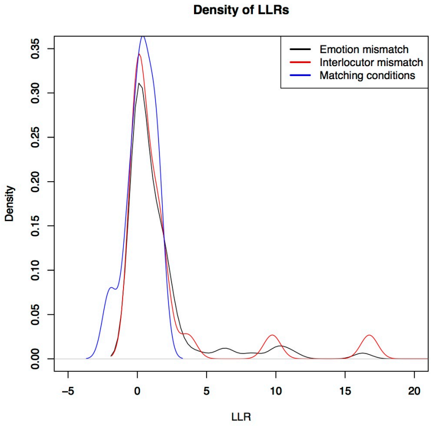

Figure 2 Distribution of LLRs for the three different comparison conditions (emotion mismatch, interlocutor mismatch, matching)

3.2.2. Effect of Emotional status

Figure 3 presents the ranges of LLR produced by various combinations of emotional states. For comparison it also shows LLRs for matching combinations. LLRs were calculated for 10 mismatch combinations and four matched combinations.

The results revealed no systematic relationship between LLRs and the types, intensity or degree of mismatch in emotional state. For instance, Combination 5 in Figure 3 (*irritated* vs *neutral_irritated*) produced the strongest LLRs with a strikingly wide variation. On the other hand, other combinations that produced larger LLRs (Combinations 4, 6, 9 and 10) varied in quality and intensity of emotion, and no commonality was found. Also Combination 2 (*irritated_angry* vs *neutral_friendly*), which was in the greatest emotional mismatch did not produce notable results. The only tendency found was that the comparisons between the samples whose emotional states were not matched tend to produce a wider range of LLRs (possibly spurious ones,) compared those with matched emotional states such as comparisons 11–14.

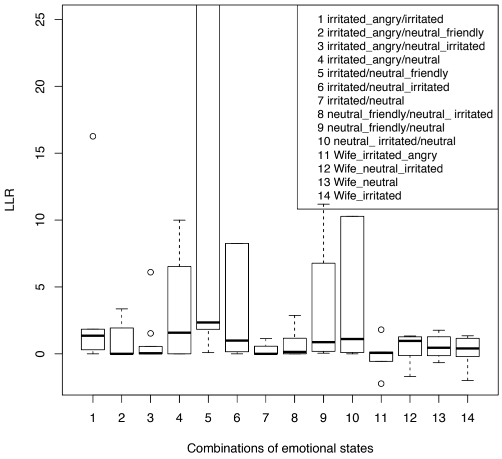

Figure 3 Comparisons of the range of LLRs based on emotional states

3.2.3. Effect of Interlocutors

The results for the interlocutor mismatch comparisons are presented in Figure 4. These did not produce as wide a range of LLRs as the emotion mismatch. Within this limited variability, however, the comparisons involving *child* as the interlocutor produced wider ranges of LLRs. There are two possible explanations for this: speaking styles and data size. The target speaker talked with his child in a clearly different speaking style. Also there were much longer recordings for the conversation between him and his wife or male friends than him and his child. So the amount of available data was very different. This effect has to be re-examined with equally sized datasets.

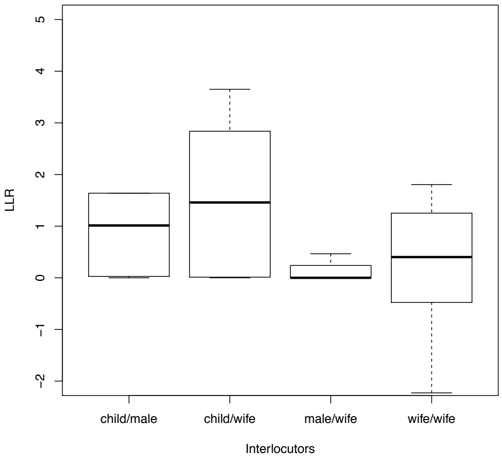

Figure 4 Comparisons of the range of LLRs based on Emotional status

4. Summary and discussion

Using authentic data from forensic casework as the testing data, this study conducted a pilot investigation on how LR-based forensic voice comparison is affected when two testing samples have a mismatch of emotional states and interlocutors.

The results revealed that forensic voice comparisons with such mismatches tend to produce wider variations of LLRs. Although no systematic relationship emerged in terms of the relationship between LLR and the types of mismatch, this finding proves the need for further study into the mechanisms at work and which effects are most significant, as well as cautioning forensic scientists working in voice comparison about the unpredictable effects of mismatched conditions.

As for future tasks, firstly the formant data for individual vowels need to be re-examined for each specific condition. As discussed above, the classification of the emotional states was far from ideal in this study. A statistically based method needs to be sought for this task. Also, comparisons with other speakers of the same variety of Australian English will be useful for separating potentially influential factors, such as ethnocultural, individual characteristics of speakers, emotional states and interlocutor.

Re-analysing against a different background population such as [20] or [21] would be also useful. While they are far from having matching conditions to the testing data used here, at least they are contemporary and spontaneous. Bernard's dataset is useful for pilot experiments and has been used in many previous studies, but the results obtained here seem to suggest that the differences in the data characteristics may have been too great. This study observed a significant number of clearly spurious LRs. These appear to be caused by the condition mismatch between the two testing recordings, but also possibly aggravated by the choice of background data. While the absolute values of LLRs were not the focus of this study, these outliers obscured the overall picture.

The number of samples used in the LR calculation needs to be controlled more. In the first instance, this study used all available data for the analysis, but the differing numbers of data points used for the LR calculation might have obscured the relationship that was the focus of this study.

Further, this study limited the analyses to formant data. Other features such as F0 and temporal features could provide other important cues.

This pilot study could not gain enough understanding to develop techniques to overcome the mismatched conditions in casework. However, it at least demonstrated that condition mismatch between testing samples can have a serious impact on LR-based voice comparison. Continuing studies are essential if we are to aspire to contribute to real life casework.

5. Acknowledgements

I would like to thank my three anonymous reviewers for their constructive comments and feedback. However, for any remaining shortcomings found in this paper, I am solely responsible.

6. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, pp. 19-41, 2000.

[2] Y. Kinoshita, "Testing Realistic Forensic Speaker Identification In Japanese: A Likelihood Ratio Based Approach Using Formants," PhD, Linguistics, The Australian National University, Canberra, 2001.

[3] D. Meuwly and A. Drygajlo, "Forensic Speaker Recognition Based on a Bayesian Framework," in *A Speaker Odyssey 2001*, Crete, Greece, 2001, pp. 145-150.

[4] P. J. Rose, T. Osanai, and Y. Kinoshita, "Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold," *The International Journal of Speech, Language and the Law* vol. 10, pp. 179-202, 12.2003 2003.

[5] G. S. Morrison, "Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/," *International Journal of Speech, Language and the Law,* vol. 15, pp. 249-266, 2008.

[6] Y. Kinoshita, S. Ishihara, and P. Rose, "Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition," *International Journal of Speech Language and the Law,* vol. 16, pp. 91-111, 2009.

[7] P. J. Rose and E. Winter, "Traditional Forensic Voice Comparison with Female Formants: Gaussian mixture model and multivariate likelihood ratio analyses," in *SST2010*, Melbourne, 2010, pp. 42-45.

[8] T. B. Alderman, "Refining the likelihood ratio approach to forensic speaker identification: The effects of non-normality in the background distribution as modelled wit the Bernard data for Australian English," Honours, School of Language Studies, The Australian National University, Canberra, 2004.

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

[10] A. Alexander, D. Dessimoz, F. Botti, and A. Drygajlo, "Aural and automatic forensic speaker recognition in mismatched conditions," *The International Journal of Speech, Language and the Law,* vol. 12, pp. 214-234, 2005.

[11] N. Brümmer and J. Du Preez, "Application independent evaluation of speaker detection " *Computer Speech and Language,* vol. 20, pp. 230-275, 2006.

[12] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice,* vol. 51, pp. 91-98, 2011.

[13] D. A. van Leeuwen and N. Brümmer, "An Introduction to Applicaiton -Independendt Evaluation of Speaker Recognition System," in *Speker Classification*. vol. 1, C. Müller, Ed., ed Berlin: Springer, 2007, pp. 330-353.

[14] D. A. van Leeuwen and N. Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," *arXiv preprint arXiv:1304.1199,* 2013.

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

[16] F. Nolan, *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press, 1983.

[17] J. Bernard, "Toward the acoustic specification of Australian English," *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung,* vol. 23, pp. 113-128, 1970.

[18] J. Harrington, F. Cox, and Z. Evansa, "An acoustic phonetic study of broad, general, and cultivated Australian English vowels," vol. Australian Journal of Linguistics, pp. 155-184, 1997.

[19] F. Cox, "The Acoustic Characteristics of /hVd/ Vowels in the Speech of some Australian Teenagers," *Australian Journal of Linguistics,* vol. 26, pp. 147-179, 2006.

[20] M. Wagner, D. Tran, R. Togneri, P. Rose, D. Powers, M. Onslow, *et al.*, "The big australian speech corpus (the big asc)," in *13th Australasian International Conference on Speech Science and Technology*, 2010, pp. 166-170.

[21] G. S. Morrison, P. Rose, and C. Zhang, "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice," *Australian Journal of Forensic Sciences,* vol. 44, pp. 155-167, 2012.