Likelihood Ratio-based Forensic Voice Comparison on L2 speakers: A Case of Hong Kong native male production of English vowels

Daniel Frost[1], **Shunichi Ishihara**[2,3]
[1]School of Language, Literature and Linguistics, The Australian National University, Australia
[2]Department of Linguistics, The Australian National University, Australia
[3]The Internet Commerce Security Laboratory, Federation University, Australia
drbfrost@live.com.au, shunichi.ishihara@anu.edu.au

Abstract

This study is a pilot research that explores the effectiveness of a likelihood ratio (LR)-based forensic voice comparison (FVC) system built on non-native speech production. More specifically, it looks at native Hong Kong Cantonese-speaking male productions of English vowels, and the extent to which FVC can work on these speakers. 15 speakers participated in the research, involving two non-contemporaneous recording sessions with six predetermined target words – “hello”, “bye”, “left”, “right”, “yes”, and “no”. Formant frequency values were measured from the trajectories of the vowels and surrounding segments. These trajectories were modelled using discrete cosine transforms for each formant (F1, F2 and F3), and the coefficient values were used as feature vectors in the LR calculations. LRs were calculated using the multivariate-kernel-density method. The results are reported along two metrics of performance, namely the log-likelihood-ratio cost and 95% credible intervals. The six best-performing word-specific outputs are presented and compared. We find that FVC can be built using L2 speech production, and the results are comparable to similar systems built on native speech.

1 Introduction

1.1 Forensic voice comparison and the likelihood-ratio framework

Forensic voice comparison (FVC) is the forensic science of comparing voices. It is most often used in legal contexts where the origin of voice samples is being debated. Typically, an FVC analysis involves the comparison of voice recordings of known origin (e.g. the suspect’s speech samples) with other voice recordings of disputed origin (e.g. the offender’s speech samples) (Rose, 2004). The FVC expert will apply statistical techniques on data extracted from speech sample evidence with the ultimate aim of assisting the trier of fact (e.g. judge(s)/jury) with their final decision. The trier of fact is faced with the task of making this decision by analysing the numerous probabilistic forms of evidence offered to them over the course of the trial. In fact, this decision is in itself a probabilistic statement, known as the posterior odds, and can be expressed mathematically as 1).

$$\frac{p(H|E)}{p(\bar{H}|E)} \quad (1)$$

In 1), $p(H|E)$ represents the probability of one hypothesis (e.g. the prosecution hypothesis – the suspect is guilty), given the various forms of evidence (e.g. DNA, fingerprint, voice, witness accounts etc.), and $p(\bar{H}|E)$ represents the probability of the alternative hypothesis (e.g. the defence hypothesis – the suspect is not guilty), given the evidence. In the context of FVC, 1) becomes:

$$\frac{p(H_{SS}|E)}{p(H_{DS}|E)} \quad (2)$$

In 2), H_{SS} represents the same-speaker hypothesis, and H_{DS} represents the different-speaker hypothesis. Before the trier of fact is able to make their decision of guilt or innocence, there may be more evidence that needs to be taken into account (e.g. DNA, fingerprint, witness etc.), and the FVC expert does not have access to this evidence (Rose, 2002, p. 57). If the FVC expert were to provide the

Daniel Frost and Shunichi Ishihara. 2015. Likelihood Ratio-based Forensic Voice Comparison on L2 speakers: A Case of Hong Kong native male production of English vowels . In *Proceedings of Australasian Language Technology Association Workshop*, pages 39−47.

trier of fact with this strength-of-hypotheses statement, they would in effect be making a statement about the suspect's guilt or innocence, which is usurping the role of the trier of fact (Aitken, 1995, p. 4; Evett, 1998; Morrison, 2009a, p. 300). This issue is resolved through the application of Bayes' Theorem, given in 3).

$$\underbrace{\frac{p(H_{SS}|E)}{p(H_{DS}\,|E)}}_{posterior\ odds} = \underbrace{\frac{p(E|H_{SS})}{p(E|H_{DS}\,)}}_{likelihood\,ratio} * \underbrace{\frac{p(H_{SS})}{p(H_{DS}\,)}}_{prior\ odds} \qquad (3)$$

By using the LR framework, the FVC expert (and the DNA expert, the fingerprint expert, etc.) is able to make an objective statement regarding the strength of the evidence, and in doing so, does not usurp the role of the trier of fact.

Put simply, the LR is the probability that some evidence would occur if an assertion is true, relative to the probability that the evidence would occur if the assertion is not true (Robertson & Vignaux, 1995, p. 17). The FVC-based LR above can be interpreted as the probability p of observing some evidence E (in FVC, this is the difference between the suspect and offender speech samples) if the same-speaker hypothesis H_{SS} is true, relative to the probability p of observing the same evidence E if the different-speaker hypothesis H_{DS} is true. For example, a calculated LR of 100 would be interpreted as follows: "the evidence is 100 times more likely to arise if the speech samples are of the same speaker, than it is if the speech samples are of different speakers". To emphasise, this is not the same as saying: "it is 100 times more likely that the speech samples are of the same speaker than of different speakers".

The process essentially involves calculating the similarity of two samples as well as the typicality of the two samples against a relevant background population. The similarity and typicality are the numerator and denominator of the LR respectively.

1.2 Non-native speakers (L2 speakers)

Since the National Institute of Standards and Technology (NIST) speaker recognition evaluations (SRE)[1] started including non-native speaker data (mostly English), a series of experiments have been carried out using L2 samples in non-forensic contexts (Durou, 1999; Kajarekar et al., 2009; Scheffer et al., 2011). However, until now, FVC research has been exclusively based on native (henceforth L1) speech production. However, crimes are obviously committed by L1 speakers and L2 speakers alike. There are therefore important practical applications to be developed from L2-based FVC research. To the best of our knowledge, this study is the first LR-based study exploring the effectiveness of an FVC system built on L2 speakers. While there have been studies that make considerations that could potentially apply to L2-based FVC, such as the selection of relevant reference samples (Morrison et al., 2012), there has not been an explicit attempt to build such a system.

The participants in this study spoke English had reasonably strong HK Cantonese "accents". Furthermore, they exhibited many tendencies of L2 speakers; stuttering, pausing to recall lexical items, using only a few set grammar patterns etc. However, we do not know how the phonetic characteristics of L2 speech affect between-speaker and within-speaker variations. One possibility is that L2 accents are not "hardwired" and therefore more fluid, potentially resulting in higher within-speaker variation; a hindrance for FVC.

1.3 Research question

Having briefly outlined the key concepts of the research, the research question is:

Can FVC work on non-native speech?

As the research question suggests, this study is exploratory in nature. We maintained tight control over many variables in order to eliminate some complexities that might arise in deeper research, in order to produce a baseline for future research. The reader should note that the aim is not to find the most effective method for L2-based FVC.

2 Research Design

Speech data were collected from 15 male speakers of Hong Kong (henceforth HK) Cantonese. We used a map task to elicit the voice samples. A map task is a simple speaking task in which the participant is provided a basic map, and the interviewer conducts a mock scenario asking for simple directions to certain places, or asks about general details of the map. The map task, conducted entirely in English, allows an interviewer to elicit large quantities of a set of words without reverting to a less natural word-list method.

[1] http://www.itl.nist.gov/iad/mig/tests/spk/

All speakers were 1) male; 2) over 18 years old; 3) HK natives; 4) identify as native speakers of HK Cantonese; and 5) completed their compulsory schooling in HK. Speakers were between 18 and 24 years of age (except one 42-year-old) and attended two non-contemporaneous recording sessions at least seven days apart (mean=12.86 days excluding an outlier of 80 days). The authors acknowledge that the number of speakers in the database is very small, though real FVC casework often involves analysis of limited amounts of data.

When performing word-specific FVC research, it is most suitable to work with common words in the English vernacular, keeping the practicalities of real casework in mind. The words given in Table 1 were chosen as the target words for both their phonetic properties and practical application. We decided to use 5 random tokens of each word to build the FVC system.

Word	GAE broad transcription	HKE broad transcription
hello	həl**əʉ**	hal**əʊ**
bye	b**ɑe**	b**aɪ**
left	l**e**ft	l**ɛ**ft
right	r**ɑe**t	r**aɪ**t
yes	**je**s	**jɛ**s
no	n**əʉ**	n**əʊ**

Table 1: Target words and broad transcriptions in GAE (General Australian English) (Harrington et al., 1997) and HKE (Hong Kong English) broad transcriptions. Target segments are in bold. Note that these transcriptions are merely representative of typical phoneme realisation.

The words in Table 1 are common English words and cover both monophthong vowel productions (stable single syllable peak with one articulatory target; "left", "yes") and diphthong vowel productions (dynamic single syllable peak with two distinct articulatory targets; "hello", "bye", "right", "no") (Cox, 2012, p. 29; Ladefoged & Disner, 2012, pp. 54-55). Diphthongs are commonly used in FVC research because they often have low within-speaker variation and high between-speaker variation. This is because a diphthong, unlike a monophthong, involves substantial movement of the formant trajectories, allowing more room for individualising information (Li & Rose, 2012, p. 202).

In our case, however, we have avoided labelling the vowels as “monophthong” or “diphthong”, because the data were extracted in a manner that captured both the formant trajectory of the vowel and the surrounding consonants and transitions where applicable. We are therefore dealing with differing levels of dynamism. Under this approach, “bye” and “right” are classed as being the most dynamic, and the least dynamic are “left”, and surprisingly, “hello”, in some speakers’ cases.

Each recording session was conducted in a soundproof recording studio using professional equipment. The recordings were made using the Audacity[2] software, preset for a 32 bit recording on a mono track at a 44.1 kHz sampling rate. They were later downsampled to 16 kHz.

The EMU Speech Database System[3] was used to analyse and annotate the recorded samples. The "forest" analysis application was used with the following settings: 3 formants to be defined (F1, F2, F3), Hamming window function with window size set to 25ms and window shift set to 5ms. The “forest” analysis performed very well in general.

2.1 Parametrisation

In order to build our FVC system, our formant trajectory portions needed to be modelled. We used a parametric curve fitting procedure that uses *discrete cosine transforms* (DCTs). The DCT method involves an estimation of a complex curve – the formant trajectories – by adding simple cosine functions together (Morrison, 2009b, p. 2389; Rose, 2013). These simple cosine functions are defined in terms of their coefficient values, which specify their amplitudes. The DCT coefficient values – from models of F1, F2, and F3 trajectories – were used as the feature vectors in the LR calculations. The durations of the trajectories were equalised because it has been shown to work well in FVC (Morrison, 2008, 2009b; Morrison & Kinoshita, 2008).

In this study, we use the term “output” to refer to the statistical and graphical result of a certain set of combinations of DCT coefficients and formants.

Figure 1 shows the modelled DCT curves (dotted lines) alongside the complex formant trajectories (solid lines) for all “bye” tokens. It is evident that higher degree DCT curves better approximate the complex formant trajectories.

[2] http://audacity.sourceforge.net/

[3] http://emu.sourceforge.net/

Table 2 shows the possible combinations of the parameters. Note that each output kept the DCT coefficient number constant across all formants in combination.

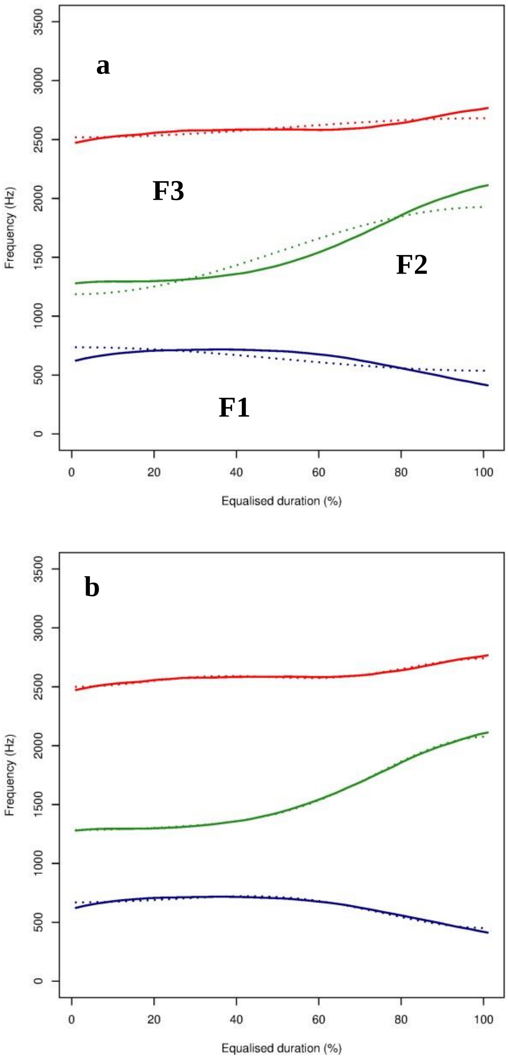

Figure 1: Solid lines represent the mean complex formant trajectories for all “bye” tokens in the dataset. The dotted lines represent 2nd degree (a) and 4th degree (b) DCT-modelled curves. X-axis = Equalised duration and Y-axis = Frequency in Hz.

3 Testing

In order to assess the performance of an FVC system, two types of comparisons, namely same-speaker (SS) and different-speaker (DS) comparisons, are necessary. In SS comparisons, two speech samples produced by the same individual are compared and evaluated with the derived LR. Given the same origin, it is expected that the derived LR is higher than 1. In DS comparisons, they are expected to receive an LR lower than 1. In total, there were 15 SS comparisons and 210 DS comparisons[4] for each target word.

Formant combination	DCT coefficients
f12, f23, f123	2, 3, 4, 5, 6, 7, 8, 9

Table 2: The multiple-formant output combinations ($6 \times 3 \times 8 = 144$ total combinations). Note that, for example, f12 represents results involving F1 and F2; f23 represents results involving F2 and F3, etc.

3.1 Multivariate-kernel-density procedure

One of the advantages of the LR framework is the ability to combine different pieces of evidence. If multiple LR values are obtained from different pieces of evidence (e.g. fingerprint, voice, DNA etc.), then these values may simply be multiplied together (added together in the logarithmic domain) to produce one LR value. This simple procedure, however, works under the assumption that the pieces of evidence are not correlated.

As explained in §2.1, DCT coefficients from models of F1, F2, and F3 trajectories were used as the feature vectors in the LR calculations. An issue here is the potential correlation between formants. The issue of correlated variables was addressed by Aitken & Lucy (2004) with their multivariate kernel density likelihood ratio (henceforth MVKD) formulae. By using a cross-validated MVKD procedure, we were able to obtain a single LR from multiple correlated features while taking the correlations into account (the statistical information for typicality is repeatedly recalculated from all samples except those speakers in comparison). The cross-validated MVKD approach has been used in many FVC studies (Ishihara & Kinoshita, 2008; Morrison, 2009b; Morrison & Kinoshita, 2008; Rose, 2013).

[4] For DS comparisons, two independent different DS comparisons are possible (e.g. (S)peaker1(R)ecording1 vs. S2R1 and S1R2 vs. S2R2) for each pair of different speakers (e.g. S1 vs. S2).

3.2 Logistic-regression calibration

When building an FVC system, raw output values may need to be calibrated before they are interpretable. The outputs of the MVKD calculations in §3.1 actually result in *scores*. Scores are logLRs in that their values indicate degrees of similarity between two speech samples having taken into account their typicality against a background popula-population (Morrison, 2013, p. 2). Logistic-regression calibration (Brümmer & du Preez, 2006) is a method which converts these output scores to interpretable logLRs by performing a linear shift (in the logarithmic scale) on the scores relative to a decision boundary.

The weights involved in the shift are calculated by using a training set of data. This involves running sets of known-origin pairs through the system to obtain scores, resulting in a training model. In an ideal situation, one would have three databases upon which to build an FVC system; the background database (used to build a model of the distribution of the acoustic feature of interest), the development database (used to calculate the weights for logistic-regression calibration and for general optimisation), and the test database (previously unused recordings that can be used to test the system – often the offender and suspect recordings) (Morrison et al., 2012). In this study, due to the limitations in the amount of data, the calibration weights were obtained using a cross-validated procedure; each derived score was referenced against every other score in the database to produce the weights. This is quite a common technique, and it has been shown to work well with MVKD-based LR outputs (Morrison, 2009b; Morrison & Kinoshita, 2008; Morrison et al., 2011).

The FoCal toolkit[5] was used for logistic-regression calibration (Brümmer & du Preez, 2006).

3.3 Metrics of performance

Evidence must be reported alongside measures of *accuracy* (also *validity*) and *precision* (also *reliability*) in order to be admitted as scientific evidence in court (Morrison, 2009a, p. 299). Accuracy refers to the "closeness of agreement between a measured quantity value and a true quantity value of a measurand" (BIPM et al., 2008, p. 21), and precision refers to the "closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions" (BIPM et al., 2008, p. 22).

Two metrics that can be used to assess output performance under this requirement are the *log-likelihood-ratio cost* (the measure of validity) (Brümmer & du Preez, 2006), and *credible intervals* (the measure of reliability) (Morrison, 2011).

Log-likelihood-ratio cost

One way of assessing validity is to find the overall correct-classification rate of the output – the equal error rate (EER). However, EER is "based on a categorical thresholding, error versus not-error, rather than a gradient strength of evidence" (Morrison, 2011, p. 93). It is not an appropriate measure of system performance as it refers to posterior probability (a question of guilt or innocence). Furthermore, these "error versus not-error" decisions are binary, unlike LRs, which are continuous; "[t]he size of a likelihood ratio indicates the strength of its support for one hypothesis over the other" (Morrison, 2011, p. 93). EER does not provide any means of assessing the strength of the LRs of an output. So, while EER can be a useful metric for the overall discriminability of a system, it is not strictly appropriate for use in FVC.

It has been argued that a more appropriate metric for assessing the validity of an output is the log-likelihood-ratio cost (henceforth C_{llr}) (Brümmer & du Preez, 2006). C_{llr} can be calculated using 4).

$$C_{llr} = \frac{1}{2}\left(\frac{1}{\mathrm{N_{H_p}}} \sum_{\mathrm{i\ for\ H_p=true}}^{\mathrm{N_{H_p}}} \log_2\left(1 + \frac{1}{\mathrm{LR_i}}\right) + \frac{1}{\mathrm{N_{H_d}}} \sum_{\mathrm{j\ for\ H_d=true}}^{\mathrm{N_{H_d}}} \log_2\left(1 + \mathrm{LR_j}\right) \right) \quad (4)$$

$\mathrm{N_{H_p}}$ and $\mathrm{N_{H_d}}$ refer to the numbers of SS and DS comparisons. $\mathrm{LR_i}$ and $\mathrm{LR_j}$ refer to the LRs derived from these SS and DS comparisons, respectively.

C_{llr} takes into account the magnitude of consistent-with-fact (and contrary-to-fact) LR values, and assigns them appropriate penalties. For example, $\log_{10}\mathrm{LR} = -5$ for an SS comparison would contribute a much heavier penalty to C_{llr} than $\log_{10}\mathrm{LR} = -0.5$ for an SS comparison. Similarly, a

[5] https://sites.google.com/site/nikobrummer/focal

correctly-classified SS comparison with $\log_{10}LR=0.5$ does not provide much support for the same-speaker hypothesis, and would therefore contribute a larger penalty than $\log_{10}LR=4$ for an SS comparison (Morrison, 2011, p. 94). For any output, an obtained C_{llr} value less than 1 implies that the output is providing a certain amount of information, and the validity gets better as C_{llr} approaches 0. The FoCal toolkit[6] was also used for calculating C_{llr} values in this study (Brümmer & du Preez, 2006).

Credible intervals

To assess reliability (precision), we used 95% credible intervals (95% *CI*). Credible intervals are "the Bayesian analogue of frequentist confidence intervals", and have the following interpretation: "we are 95% certain that the true value of the parameter we wish to estimate lies within the 95% credible interval" (Morrison, 2011, p. 95). In this study, uniform prior odds are assumed, so the actual calculations are identical to frequentist confidence intervals. It is also important to note that as there were only two recordings of each speaker, 95% *CI* values can only be estimated from the DS comparisons.

4 Results

Table 3 shows the best-performing outputs for each target word in terms of C_{llr}.

word	C_{llr}	**formant combination**	**DCT coefficients**	**95% CI**
Bye	0.158	23	5	9.996
Right	0.271	123	2	7.272
No	0.318	123	2	3.472
Left	0.342	123	2	4.249
Hello	0.392	123	2	3.518
Yes	0.527	23	5	4.232

Table 3: Best-performing outputs for each target word by C_{llr}.

Table 3 shows that "bye" performed best in terms of C_{llr}, and "yes" was the worst by the same measure. However, on closer inspection we see that the 95% *CI* for "bye" is poor in comparison to the other words. This is not a coincidence; "bye" consistently performed the best in terms of C_{llr} even with other combinations of the parameters, while performing the worst in terms of 95% *CI*.

A Pearson correlation test shows a negative correlation between the C_{llr} and 95% *CI* values (= -0.700; p < 0.0001) across all words. This is actually to be expected; Morrison (2011)) notes that one would ideally hope for low values for both metrics, but in practice, this is not often the case. It is clear that there is a trade-off when it comes to assessing the performance of the outputs.

When comparing the typical trajectories of the vowels in these words, it is noticeable that performance, in terms of C_{llr}, roughly corresponds to the level of dynamism of the trajectories. 2nd, 3rd, 4th, and 5th degree DCT-modelled curves tended to perform the best.

Presented in Figure 2 are the Tippett plots for the best-performing outputs of each word. Tippett plots show the cumulative distribution of $\log_{10}LRs$ for SS and DS comparisons. As stated earlier, in a good output we expect most SS comparisons to produce $\log_{10}LRs > 0$, and most DS comparisons to produce $\log_{10}LRs < 0$. The counter-factual LRs (circled in Figure 2a as an example) that are "penalised" by C_{llr} (and their strength) become clear when inspecting a Tippett plot. The EER is also made clear in a Tippett plot; it is the crossing point of the SS and DS lines (indicated by the arrow in Figure 2e as an example). 95% *CI* bands (grey dotted curves) are also included in the Tippett plots given in Figure 2 for the DS comparison curves.

As can be seen in Figure 2, in all outputs, the DS LRs achieve greater values compared to the SS LRs; the DS curves are less steep than the SS curves. This is partly due to the number of DS comparisons (210) in each output outnumbering the number of SS comparisons (15). Also, when counter-factual, the DS comparisons tend to be more counter-factual than SS comparisons (except "yes" SS comparisons).

It is immediately obvious that "bye" is the highest performer; it achieves the greatest SS and DS values of all the outputs (values furthest away from $\log_{10}LR = 0$) and it has 100% correct discrimination for SS comparisons. It does produce misleading DS LRs, but the strength of these LRs is comparable with the other outputs. "No" also achieves 100% correct discrimination for SS comparisons, and "right" and "left" come very close to doing so.

[6] https://sites.google.com/site/nikobrummer/focal

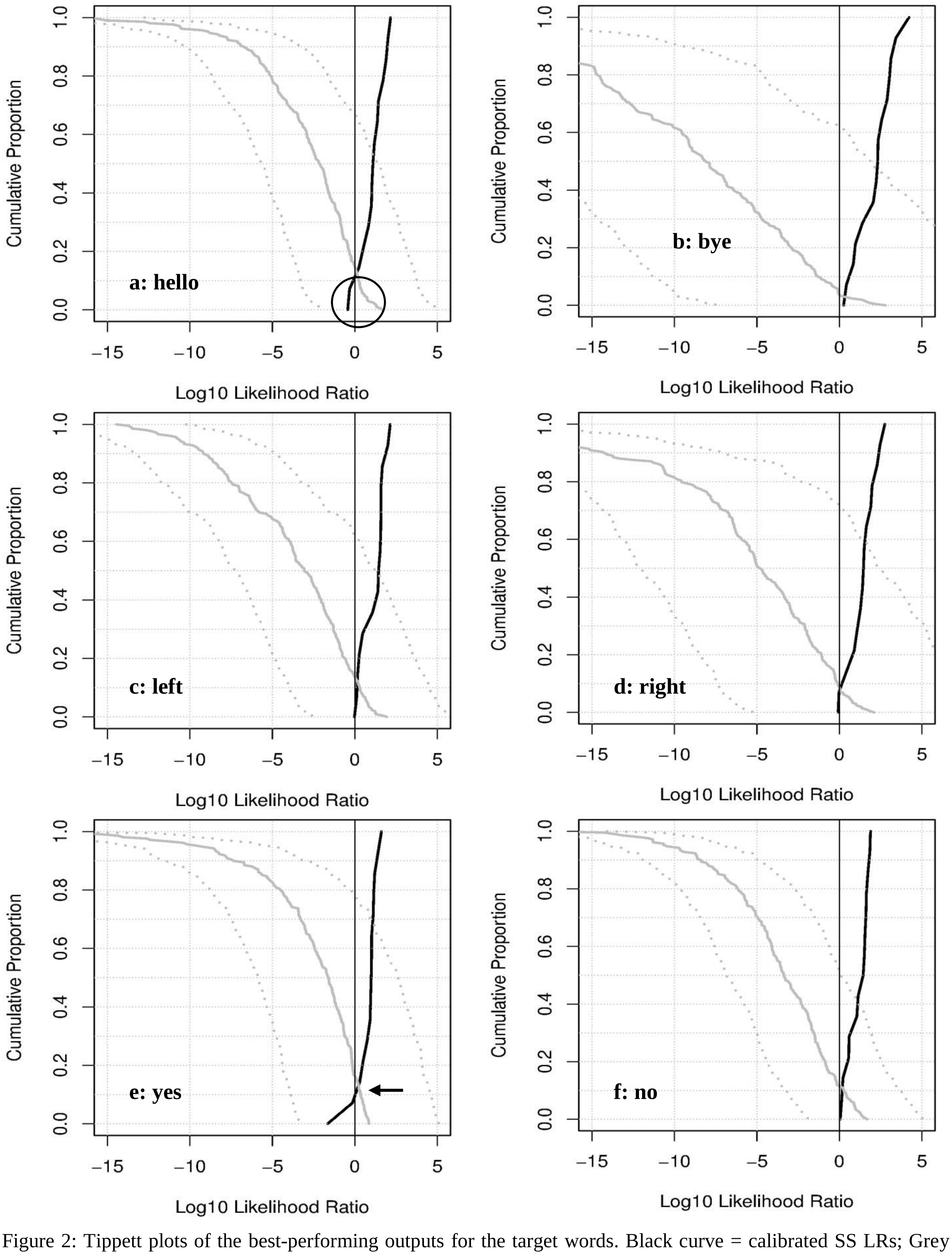

Figure 2: Tippett plots of the best-performing outputs for the target words. Black curve = calibrated SS LRs; Grey solid curve = calibrated DS LRs. Dotted grey curves = 95% *CI* band. The circle in Panel A indicates the counter-factual LRs and the arrow in Panel E indicates the EER.

5 Discussion

For the participants in this study, phonetic realisation varied greatly between speakers, and speakers were generally consistent internally. Table 4 and Table 5 show various phonetic realisations for “bye” and “hello” respectively. The effect of this variation is seen in the overall performance of our system; our DS comparisons tended to perform very well.

bye			
consonant	start target	length	end target
b p	a ɑ æ ɐ	unmarked ˘ ˑ ː	i ɪ e ə

Table 4: Various phonetic variations seen in the production of “bye”. (Not all combinations were realised– this is a list of articulations that appeared in the given positions.)

hello				
consonant	vowel	consonant	target 1	target 2
h	ɛ ə ɐ a	l ˡ ɾ Ø	ə ɜ ɛ o	ʊ u

Table 5: Various phonetic variations seen in the production of “hello”. Another common final vowel was [oː].

While our research aim makes no mention of a comparison of our L2-based FVC system with similar traditional L1-based FVC systems, it is still an issue of particular interest. While it is not theoretically appropriate to directly compare C_{llr} values between systems unless the experimental settings are identical, doing so can provide a rough comparison of two systems. Morrison (2009b) looked at parametric representations (DCTs and polynomials) of the formant trajectories of five Australian English diphthongs, namely /aɪ/, /eɪ/, /oʊ/, /aʊ/, /ɔɪ/ (/aɪ/ corresponds to the /aɪ/ in this study, and /oʊ/ corresponds to the /əʊ/ in this study) from 27 Australian males. The best /aɪ/ output achieved a C_{llr} of 0.156, compared to 0.158 (“bye”) and 0.271 (“right”) in this study, and the best /oʊ/ (/əʊ/) output achieved 0.129, compared to 0.318 (“no”) and 0.392 (“hello”) in this study. We can see that the performance of the diphthong-specific outputs is quite comparable to the equivalent outputs in this study. This implies that L2-based FVC systems have no major shortcomings.

6 Conclusion

This study was the first to build an LR-based FVC system on L2 speech production, motivated by the relative prevalence of crimes involving L2 speakers. 15 native HK Cantonese-speaking males participated in the research. Six common words were targeted, and DCT-modelled parametric curves were fitted to the formant trajectories of the six target words. The coefficient values of the DCT-modelled curves were used as feature vectors in the LR calculations. The MVKD procedure (Aitken & Lucy, 2004) was used to produce LRs for each word. We used logistic-regression calibration (Brümmer & du Preez, 2006) to calibrate the outputs of the MVKD procedure.

Each output was evaluated with two metrics; the log-likelihood-ratio cost (C_{llr}) measured validity, and credible intervals (95% *CI*) measured reliability. We found that the words with more dynamic formant trajectories tended to perform best, and outputs involving F1, F2 and F3 performed better than outputs involving just F1 and F2, or F2 and F3. 2nd, 3rd, 4th, and 5th degree DCT-modelled curves tended to produce the best outputs.

In terms of the research question – whether or not FVC can be performed on L2 speech – we have clearly demonstrated that FVC can, and does, work on L2 speech. Further, we achieved results comparable to traditional L1-based FVC systems, which is certainly promising for the prospects of the field.

Acknowledgments

The authors greatly appreciate the very detailed comments of the three anonymous reviewers. This study is part of the first author’s honours thesis at the Australian National University.

References

Aitken, C. G. (1995). *Statistics and the evaluation of evidence for forensic scientists*. Chichester: J. Wiley.

Aitken, C. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 53*(1), 109-122.

BIPM, I., IFCC, I., IUPAC, I., & ISO, O. (2008). Evaluation of measurement data—guide for the expression of uncertainty in measurement. JCGM 100: 2008.

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language, 20*(2), 230-275.

Cox, F. (2012). *Australian English pronunciation and transcription*. Cambridge: Cambridge University Press.

Durou, G. (1999). Multilingual text-independent speaker identification. *Proceedings of the Multi-Lingual Interoperability in Speech Technology*, 115-118.

Evett, I. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice, 38*(3), 198-202.

Harrington, J., Cox, F., & Evans, Z. (1997). An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Australian Journal of Linguistics, 17*(2), 155-184.

Ishihara, S., & Kinoshita, Y. (2008). How many do we need? Exploration of the population size effect on the performance of forensic speaker classification. *Proceedings of the Interspeech 2008*, 1941-1944.

Kajarekar, S. S., Scheffer, N., Graciarena, M., Shriberg, E., Stolcke, A., Ferrer, L., & Bocklet, T. (2009). The SRI NIST 2008 speaker recognition evaluation system. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, 4205–4208.

Ladefoged, P., & Disner, S. F. (2012). *Vowels and Consonants*. Chichester: Wiley.

Li, J., & Rose, P. (2012). Likelihood ratio-based forensic voice comparison with F-pattern and tonal F0 from the Cantonese /ɔy/ diphthong. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, 201-204.

Morrison, G. S. (2008). 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, 15*, 247-264.

Morrison, G. S. (2009a). Forensic voice comparison and the paradigm shift. *Science & Justice, 49*(4), 298-308.

Morrison, G. S. (2009b). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America, 125*, 2387-2397.

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice, 51*(3), 91-98.

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences, 45*(2), 173-197.

Morrison, G. S., & Kinoshita, Y. (2008). Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English /o/ Formant Trajectories. *Proceedings of the Interspeech 2008*, 1501-1504.

Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. *Proceedings of the Odyssey 2012*, 62-77.

Morrison, G. S., Zhang, C., & Rose, P. (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic science international, 208*(1), 59-65.

Robertson, B., & Vignaux, G. A. (1995). *Interpreting evidence*. Chichester: Wiley.

Rose, P. (2002). *Forensic speaker identification*. London & New York: Taylor & Francis Forensic Science Series.

Rose, P. (2004). Technical forensic speaker identification from a Bayesian linguist's perspective. *Proceedings of the Odyssey 2004*, 3-10.

Rose, P. (2013). More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *International Journal of Speech Language and the Law, 20*(1), 77-116.

Scheffer, N., Ferrer, L., Graciarena, M., Kajarekar, S., Shriberg, E., & Stolcke, A. (2011). The SRI NIST 2010 speaker recognition evaluation system. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, 5292-5295.