THE EFFECT OF SPEAKER SAMPLING IN LIKELIHOOD RATIO BASED FORENSIC VOICE COMPARISON

Bruce Xiao Wang, Vincent Hughes and Paul Foulkes

Abstract

Within the field of forensic voice comparison (FVC), there is growing pressure for experts to demonstrate the validity and reliability of the conclusions they reach in casework. One benefit of a fully data-driven approach that utilises databases of speakers to compute numerical likelihood ratios (LRs) is that it is possible to estimate validity and reliability empirically. However, little is known about the stability of LR output as a function of the specific speakers sampled for use in the training, test, and reference data sets. The present study addresses this issue using two large sets of formant data: Cantonese sentence final particle /a/ and British English filled pauses UM. Experiments were replicated 100 times varying the (1) training, test and reference speakers, (2) training speakers only, (3) test speakers only, and (4) reference speakers only. The results show that varying the speakers in all three sets has the greatest effect on system stability for both the Cantonese and English variables, with the C_{llr} varying from 0.60 to 0.97 for /a/ and 0.32 to 1.33 for UM. However, this variability is primarily due to the effects of uncertainty in the test set. Varying only the training speaker speakers has the least effect on system stability for /a/ (C_{llr} range: 0.76 to 0.88), while varying reference speakers has the smallest effect for UM (C_{llr} range: 0.40 to 0.54). The results indicate that in LR-based FVC it is important to assess the stability of system as a function of the samples of speakers used (C_{llr} range) rather than just reporting a single C_{llr} value based on one configuration of speakers in each set. The study contributes to the general debate on reporting uncertainty in LR computation.

Key words: forensic voice comparison, Likelihood ratio, system stability, English filled pause, Cantonese sentence final particle.

1 Introduction

Forensic voice comparison (FVC) involves the comparison of two or more speech samples in the context of a legal case; typically one of an unknown offender, and the other of a known suspect typically recorded during a police interview, e.g. in the UK (Home Office 2003) or through wiretaps, e.g. in Germany and China (Liu 2006). In the last two decades, the likelihood ratio (LR) framework has been employed in more and more FVC studies (e.g. Morrison 2008; Zhang, Morrison and Thiruvaran 2011; Hughes, Wood and Foulkes 2016), and is now widely accepted, at least in principle, as the "logically and legally correct" (Rose and Morrison 2009: 143) approach for the evaluation of comparison evidence across forensic sciences (Robertson and Vignaux 1995). The LR is a measure of the strength of evidence under the competing propositions of the prosecution and defence, which takes into account the similarity between the suspect and offender samples, as well as their typicality in terms of a relevant population (see Hughes and Foulkes 2015 for more details about relevant reference populations). For example, if the samples are very similar to one another, the strength of evidence is high when the typicality is low (i.e. it is not likely to find a similar speech sample among the relevant population), while the strength of evidence is low when the typicality is high (i.e. it is likely to find a similar sample from the relevant population).

Data-driven LR-based FVC relies on corpora of speakers to estimate empirically the strength of the voice evidence. It is essential to do empirical testing of validity and reliability of a system (i.e. how well the system is able to separate/discriminate between same- and different-speaker pairs and the stability of the output), in order for the end-user to be able to contextualise the LR conclusion provided. It is worth noting that the term *system* here refers to the method, speech features and corpus that are used for LR computation (Morrison 2013), rather than a much narrower meaning that used to describe automatic speaker recognition software. The system is normally evaluated by using three sets of data, namely training, test and reference data sets. The training data is used to train a model which is applied to the test data for system calibration, and the reference data is used for the evaluation of typicality. However, there are inevitable uncertainties in data-driven analysis. That is, the LRs and overall system performance will reflect the specific choices and groupings of data. Specifically, this may include the speakers used for training and testing the system, the variables used as input, the number and choice of tokens, and the methods used for extracting acoustic data etc. The underlying uncertainty could also be introduced from the data generating process such as sampling from the relevant population, and the statistical modelling technique employed (Morrison 2016). Recognition of these facts has generated debate on how to calculate and report precision or uncertainty in LR calculations, reflecting "imperfect data sources and imperfect knowledge" (Curran 2016: 382; see also e.g. Morrison and Enzinger 2016).

Many previous studies have explored the performance of linguistic features such as individual vowels and phonetic sequences using LR-based testing (e.g. Morrison 2009; Zhang et al. 2011; Rose and Wang 2016). Typically, a group of speakers (often 60) is selected, split equally into training, test and reference speakers (e.g. 20-20-20), before running the analysis. Some studies have explored system stability by taking sociolinguistic factors into consideration, e.g. age, social class (Hughes and Foulkes 2015) and accent (Hughes and Foulkes 2017), which shows that the system yields better performance when using speakers with matched sociolinguistic factors (i.e. matched accent/age group/social class). Other studies have looked into the effect of the number of speakers on system performance, i.e. how many speakers are enough to achieve stable and accurate LR output? (e.g. Ishihara and Kinoshita 2008, 2014; Hughes 2017). Ishihara and Kinoshita (2014) found that the system starts to yield stable performance when the number of reference speakers reaches 30, and the system performance is close to optimum when there are 70 reference speakers, while Hughes (2017) found that the system starts to yield stable performance when more than 20 speakers are used for each of the training, test and reference data sets.

However, no previous studies have explored 'who' rather than 'how many' speakers should be used in LR-based FVC. One key question is whether system performance is consistent if we use different samples of speakers from a single relevant population (e.g. speakers with similar social class, accent, education etc.) and replicate experiments multiple times. The current study builds on Hughes (2017) to explore the stability of LR output according to the specific speakers used (rather than size) for training, test, and reference data sets. The current study uses two segmental variables: the Cantonese sentence final particle /a/ and the filled pause UM from Standard Southern British English (SSBE). Four experiments were carried out to explore the following questions.

1. To what extent does system performance vary if different sets of training, test and reference speakers (of the same size) sampled from the same relevant population are used?

2. Is the training, test or reference data more sensitive to speaker sampling?

3. What is the feasibility of using the same system for multiple cases?

Section 2 gives more details about the corpora, variables and data processing. However, only Cantonese data processing is explained in detail since it was collected specifically for the present study, Existing data were used for UM. The data collection and processing are described in from Hughes et al. (2016). Section 3 explains the experiment procedures and results are presented in section 4. Section 5 discusses the experimental results with regard to the three research questions and compares the results with previous studies, and a conclusion is given in section 6.

2 Method

2.1 Corpora

Two corpora were used in the present study. The first is the IARPA Babel Cantonese language pack (Andrus et al. 2016), which contains approximately 215 hours of Cantonese conversational and scripted telephone speech of speakers from Guangdong province in mainland China. The corpus was designed for training speech recognition technologies. Therefore, it is not an ideal corpus for FVC studies, because there was only one recording session per speaker. As such it does not fit with typical forensic conditions involving two samples recorded with some time gap between them. Using contemporaneous speech data of this kind is expected to overestimate the accuracy and stability of the overall FVC system (Enzinger and Morrison 2012) relative to real casework. Sociolinguistic factors are also not controlled in this corpus, beyond language (Cantonese) and biological sex (male). However, this corpus is forensically realistic in terms of channel since the conversations were recorded using different telephones in different environments (e.g. indoor, outdoor, recording/transmission-channel mismatch). In principle, forensic recordings could be made in any situation by any recording devices. All the audio files were sampled at a rate of 8000Hz, meaning that only information up to 4000Hz was available for analysis, and in 8-bit a-law encoded sphere format. Due to poor transmission and different quality of telephone voice recorders, the third formant of vowels proved extremely difficult to measure. As a result, only the first two formants were used in the current study.

The second corpus is the Dynamic Variability in Speech corpus (DyViS; Nolan, McDougall, de Jong, and Hudson 2009). DyViS was designed for forensic phonetic research. It contains 100 Standard Southern British English (SSBE) male speakers aged between 18 and 25. The data used for the current study were extracted from Task 1 and Task 2. Task 1 involves a mock police interview where speakers assumed to be the suspects, and were asked to answer questions based on a map given on a screen, while avoiding incriminating information. The purpose of this task is to obtain “spontaneous speech in a situation of ‘cognitive conflict’, where speakers were made to lie” (Nolan et al. 2009: 41). Task 2 involves a telephone call with a ‘accomplice’, where the researcher requests a debriefing from the subject about the mock police interview. The DyViS corpus is less forensic realistic as all the recordings are high studio quality.

2.2 Variables

Filled Pauses in SSBE

Filled pauses commonly occur in spontaneous speech and, amongst other things, fill gaps between utterances. They have been shown to be good speaker discriminants, outperforming lexical vowels in FVC tasks (Hughes et al. 2016). There are a number of potential explanations for this. Firstly, filled pauses are often abutted by silence on one or both flanks and are thus less influenced by coarticulation. Secondly, they have a high frequency of occurrence among speakers in most forms of spontaneous speech (Tschäpe et al. 2005). It is therefore likely that they will be available in most spontaneous speech samples. Thirdly, filled pauses usually have longer duration than lexical vowels, which gives longer and more stable formant trajectories making the segmentation and acoustic measurement easier to conduct (Shriberg 2001, Hughes et al. 2016). Given their impressive speaker discrimination performance in previous studies, it is of interest to assess the stability of the performance of filled pauses using different sets of training, test and reference speakers from a single relevant population. The two most common filled pauses are UH (err) and UM (erm). However, only UM was used in current study because it yielded a better speaker-discriminatory performance in Hughes et al. (2016).

Cantonese Sentence Final Particle /a/

Cantonese sentence final particles are bound forms attached in sentence final position (Law 2002). Functionally, they are often said to be the equivalent of intonation in English (Wakefield 2011). These particles are potentially good variables for FVC because they occur frequently in daily usage (Leung 2009). The number of different sentence final particles in Cantonese ranges from 30 (Kwok 1984) to 95 depending on how one counts them (e.g. /a/, /za/, /ma/, /la/) (Law 2002). In the current study we focus on /a/ '啊 ah' as it is one of the most common in Cantonese (Sybesma and Li 2006:1774). The final syllable lengthening provides the duration required for "more closely approximate canonical formant values" (Linblom 1963). Similar to filled pauses, the sentence final particle also has longer duration than inter-syllabic /a/ making the segmentation process easier. A waveform and spectrogram representation of an example of a sentence final particle /a/ is given in Figure 1. The transcript in the Praat TextGrid (version 6.0.36; Boersman and Weenink, 2017) is Romanised Cantonese pronunciation and the following number indicates tonal information. *Dei6fong1* means "place/area", while *a1* is the sentence final particle.

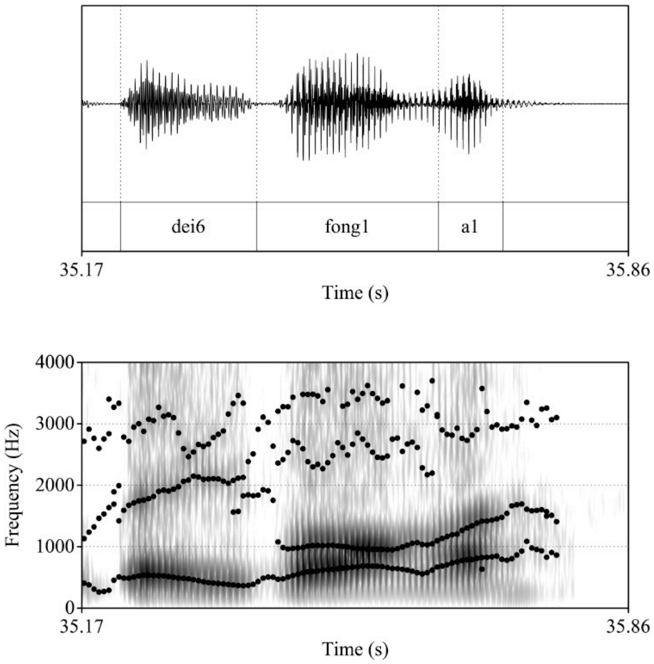

Figure 1. Example of Cantonese sentence final particle in the phrase *dei6 fong1 a1*.

2.3 Data processing

Existing filled pause data were available from Hughes et al. (2016), consisting of quadratic polynomial coefficients extracted from the first three formants of the vowel portion of UM tokens. Tokens were manually marked using Praat (version 6.0.36; Boersma and Weenink, 2017) TextGrid, and multiple measurements from across the duration of the first three formants were taken using a Praat script. After that, quadratic polynomial curves were fitted to raw formant data, and the coefficients were saved and processed for LR computation (see Hughes et al. 2016 for details). A total of 73 speakers with an average of 14 tokens per speaker per session were available.

The following sections describe in detail the extraction and processing procedures for the Cantonese data.

2.3.1 Raw data segmentation

Tokens of Cantonese /a/ were manually segmented and labelled using a TextGrid in Praat. Figure 2 shows an example of segmented token. The boundaries were placed at the onset and offset of the full vocalic portion of each token. The onset of /a/ was marked by the start of regular periodicity of the full vocalic portion, and the offset was marked by the end of periodicity. Tokens were discarded when the TextGrid boundaries could not be confidently placed due to poor recording quality and transmission issues.

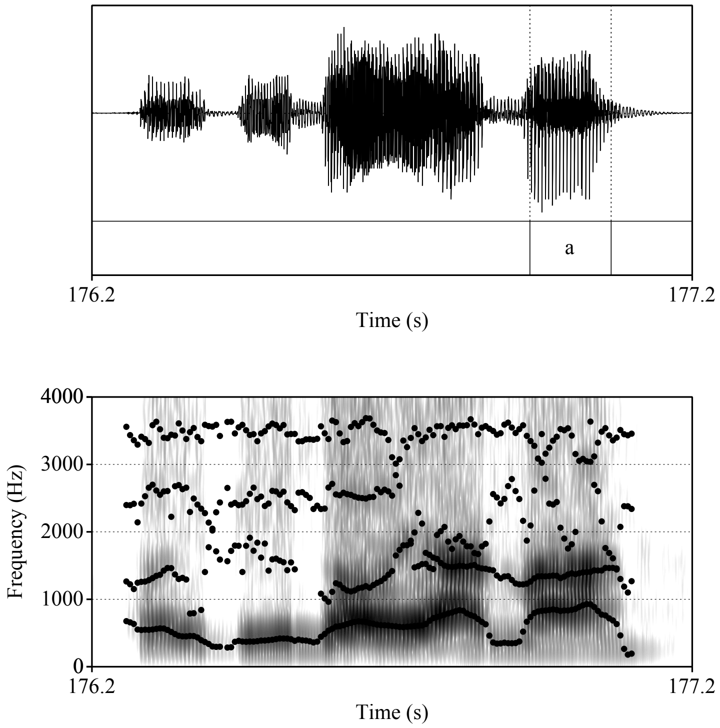

Figure 2: Segmented Cantonese sentence final particle /a/.

Two Praat scripts were then used to extract the first two formants of all the segmented tokens. As already noted, F3 was extremely weak meaning it was generally difficult to track reliably (see Figures 1 and 2 for illustration). The first script (Lennes 2003a) extracted all the segmented tokens into a separate sound file and saved it to a specified directory. The second script (Lennes 2003b) took all the individual sound files and extracted the first two formants at time-normalised +10% steps across the vocalic portion of each token (McDougall 2004, 2006) using the *To Formant (burg) ...* function in Praat. The default formatting setting was 4 formants below 4000Hz. A series of heuristics was applied to correct formant measurement errors following procedures in Hughes and Foulkes (2015). Ranges of 200-900 Hz and 1000-2000 Hz were chosen for F1 and F2. Tokens with values outside this range were removed, because they are outside reasonable frequency ranges of F1 and F2, and thus classed as obvious measurement errors. Statistical outliers were then identified by calculating the pooled mean across all speakers for each measurement point. z-scores were then calculated at each +10% step. Values of ±3.29 standard deviations greater than the mean were removed. In order to preserve as many tokens as possible, missing interval (10%) values were inspected visually and replaced by the mean of two adjacent measurements. However, the whole token was removed where the first and/or last measurements were missing.

2.3.2 Parametric curve fitting and feature extraction

Quadratic polynomial curves were fitted over the nine measurements of F1 and F2 for /a/. The polynomial coefficients capture the dynamic properties of the trajectories of formants, which reduce the dimensionality of the data set and improve discrimination performance (Hughes et al. 2016, McDougall 2006). Figure 3 shows an example of quadratic fitting for /a/ from speaker 54. The circle markers are the raw F1 and F2 values, while the lines are polynomial curves fitted to the raw F1 (darker lines) and F2 (lighter lines) values. The reason to use quadratic polynomial curves for /a/ is because its trajectory is essentially linear with no more than one turning point.

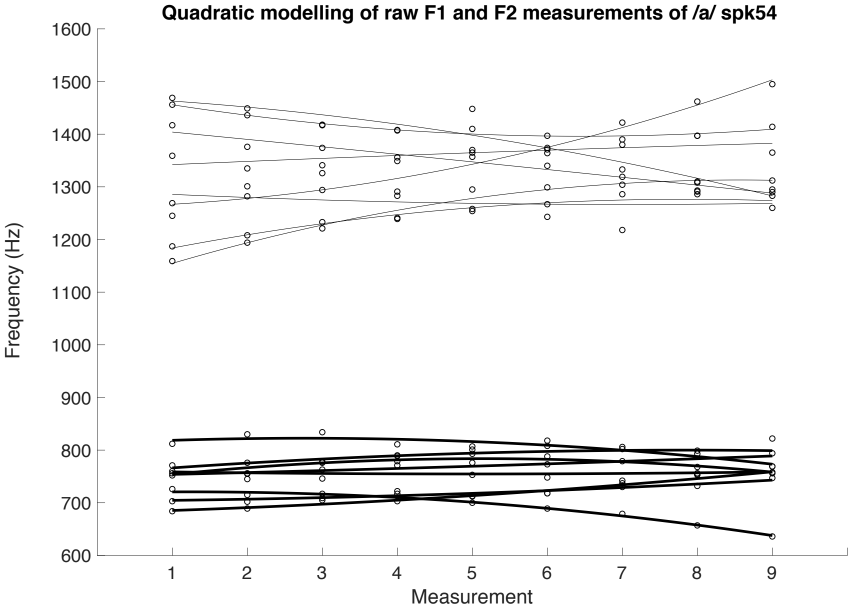

Figure 3: Quadratic curve fitting to /a/ of one speaker.

Each token was plotted and fitted with corresponding polynomial curves. The data were inspected visually, and obvious measurement errors removed based on large residuals. Polynomials were re-fitted once these measurement errors had been removed. The final data set contained 155 speakers with an average of 14 tokens per speaker in total. The input data for computing LRs consisted of three quadratic polynomial coefficients per formant which capture information about absolute frequency and the shape of the trajectory. This is a means of representing the formant dynamics with a smaller number of coefficients.

2.3.3 LR computation

LR computation involves two stages: *feature-to-score* conversion and *score-to-LLR* mapping (Morrison 2013). For Cantonese /a/, data for each speaker was divided in half with the first half acting as the suspect sample and the second half acting as the offender sample. Given that the UM data came from two recording sessions, the first session was used as the suspect sample, while the second was used as the offender sample.

At the *feature-to-score* stage, same- (SS) and different-speaker (DS) pairs in each of the training and test sets are compared using the multivariate kernel density formula (MVKD, Aitken and Lucy 2004) to produce a series of scores. This involves assessing the similarity between the suspect and offender data, and evaluating typicality using data from the reference set. MVKD uses a normal distribution to model suspect data, while the reference data are modelled using kernel density estimation based on equally weighted Gaussians for each reference speaker (in this way the estimation of typicality for MVKD is speaker-dependent; Morrison 2011). At the *score-to-LLR* stage, the training scores were then applied to train a logistic regression model (Morrison 2011). The model coefficients were then applied to the test scores to produce calibrated Log LRs (LLRs). The purpose of this is to optimise the system and “ameliorate what would otherwise be very misleading results” (Grigoras et al. 2013).

For Cantonese /a/, 30 speakers were used in each of the training, test and reference sets. Each DS pair produced a single score, which resulted in 30 SS and 435 DS scores for each of the training and test sets. For UM, only 20 speakers were used in each set to allow for different samples of speaker to be used (given that the total number of speakers available was only 73). 20 SS and 190 DS scores were produced for training and test data set respectively. System performance was evaluated using equal error rate (EER) and the log LR cost function (C_{llr}, Brümmer and du Preez 2006). EER represents the absolute proportion of trials that produce contrary-to-fact LRs (i.*e. errors*), reflecting the point at which the proportion of false hits and misses is equal. C_{llr} is gradient measure of the goodness of the LRs, accounting for the magnitude of errors rather than the absolute proportion (i.e. a large magnitude error is much more problematic for a system than a small magnitude error). In both cases, the lower the values the better the performance. For C_{llr} a value of less than one indicates that the system captures some useful information. Values of greater than one reflects very poor performance.

3 Experiments

Four experiments were conducted to explore different aspects of system testing. 100 replications of each experiment were conducted varying the speakers assigned to one or more of the training, test and reference sets. Ideally, a stable system would yield a consistent C_{llr} and EER irrespective of the make-up of the training, test, or reference sets. A larger range for the C_{llr} and EER indicates the system is highly sensitive to the make-ups of the sets.

- Experiment 1: varying all speakers

Speakers were randomly selected and assigned to the three data groups (training, test and reference) in each replication. Experiment 1 intends to mimic what happens in LR-based FVC research which often uses a single configuration of speakers in each set.

The following three experiments vary the make-up of each data set separately, to assess the relative contribution of the training, test and reference speakers to a system’s stability.

- Experiment 2: varying test speakers

Only test speakers were varied in each replication. The same set of training and reference speakers were used through 100 replications. In this way the speakers that would constitute the ‘system’ in FVC (i.e. the training and reference data) remain fixed throughout the replications. This allows us to assess the effect of only varying the speakers used to test the system.

- Experiment 3: varying reference speakers

Only reference speakers were varied in each replication, while training and test were fixed. Experiment 3 aims to explore whether a random selection of speakers from a relevant population adequately represents the population, and the sensitivity of system to the reference data.

- Experiment 4: varying training speakers

Only training speakers were varied in each replication, with test and reference sets fixed. Experiment 4 aims to explore the sensitivity of training data to different speakers, i.e. to assess the sensitivity of the system performance to different sets of the selection of speakers chosen to represent a matched population.

The experiments were carried out in R (R core team 2018) using a LR calculation and testing in FVC package (Lo 2018), which is an adaptation of the MATLAB implementation (Morrison 2007) of Aitken and Lucy's (2004) MVKD formula. The R script randomly selects speakers based on pre-defined randomisation rules for Experiments 1 to 4, runs comparison and calibration, and saves the results into a list. Each experiment was replicated 100 times with different configurations of training, test and reference speakers. Details of the experimental results are discussed below.

4 Results

4.1 Experiment 1: Varying all speakers

Figure 4 shows system performance when varying speakers in all three data sets. The boxplots show the variation in C_{llr} (left panel) and EER (right panel) for Cantonese /a/ and SSBE UM. Varying test, training, and reference speakers causes system performance to vary to different extents for the two variables. Over the 100 replications, the overall range of C_{llr} for /a/ is 0.37, while the interquartile range is 0.07. All of the C_{llr}s for /a/ are lower than 1, which indicates that the system is capturing some useful information in each replication. However, the variability in C_{llr} indicates that the system stability is sensitive to different composition of speakers in training, test, and reference data.

The C_{llr} for UM ranges from 0.32 to 1.33, and the interquartile range is 0.16. 75% of the C_{llr}s are lower than 0.64, while 50% of the C_{llr}s are between 0.48 and 0.64. There are seven statistical outliers among these results and three of them are larger than one, meaning that the system is not capturing any useful information in those three replications. Four other replications had C_{llr}s larger than 0.9, which also indicates a fairly poor performance. In general, UM yielded less stable system performance than the Cantonese /a/, although overall UM is a better speaker discriminant with most of the Cllrs for than those of /a/ across the 100 replications. It is also noted that none of the replications of UM yielded a better C_{llr} than that in Hughes et. al (2016).

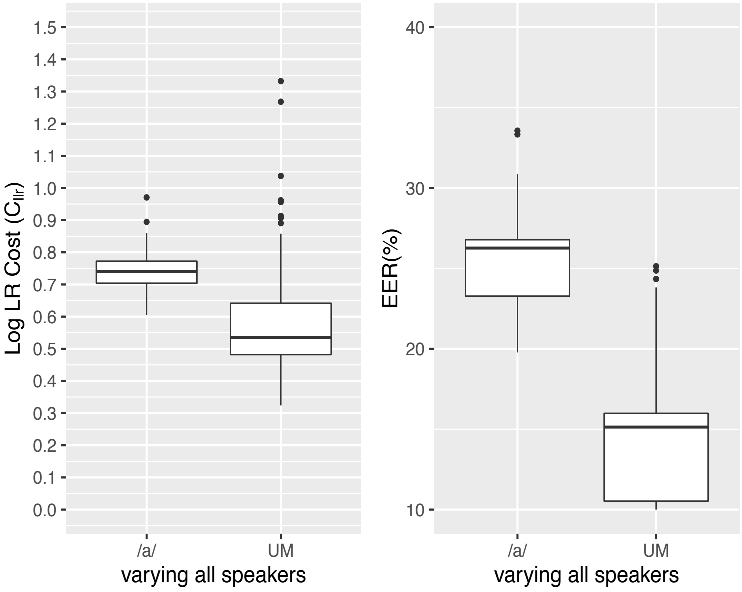

Figure 4: Boxplots of C_{llr} and EER of /a/ and UM by varying training, test and reference speakers in each replication.

	C_{llr}		EER(%)	
	/a/	UM	/a/	UM
Minimum	0.60	0.32	19.8	10.0
1st quartile	0.70	0.48	23.3	10.5
Median	0.74	0.54	26.3	15.1
3rd quartile	0.77	0.64	26.8	16.0
Maximum	0.97	1.33	33.6	25.1
Interquartile range	0.07	0.16	3.5	5.5
Overall range	0.37	1.01	13.8	15.1

Table 1: Minimum, maximum, median, interquartile range, overall range, first and third quartiles of C_{llr}s and EERs of /a/ and UM in experiment 1.

EER shows a similar pattern to C_{llr} (Figure 4). The EER for /a/ varies from 19.8% to 33.6%, with a median of 26.3% and an interquartile range of 3.5%. The EERs for UM range from 10% to 25.1% (OR = 15.1%) over the 100 replications. The median is 15.13% and the interquartile range is 5.46%. As with C_{llr}, the EER results indicate that UM produces less stable performance compared with Cantonese /a/, although it is a better speaker discriminant; over 75% of EERs for UM are lower than the EERs for /a/.

4.2 Experiment 2: Varying test speakers

In experiment 2, training and reference speakers were fixed, while different samples of test speakers were used in each replication. The summary results are shown in Table 2. As in experiment 1, Figure 5 shows the two variables tested here are affected differently by varying the test speakers. The C_{llr} for Cantonese /a/ ranges from 0.58 to 0.86 across the 100 replications. The median is 0.74 while the interquartile range is 0.08, indicating that 50% of C_{llr}s are within a small range. Similar to experiment 1, all the C_{llr}s of /a/ are lower than one, indicating that some speaker discriminatory information is being captured. The minimum and maximum C_{llr}s for UM are 0.33 and 0.94, which represents considerably more variability than that found for /a/. The median for UM, however, is 0.67, indicating the overall speaker discriminatory power for UM is greater than for /a/.

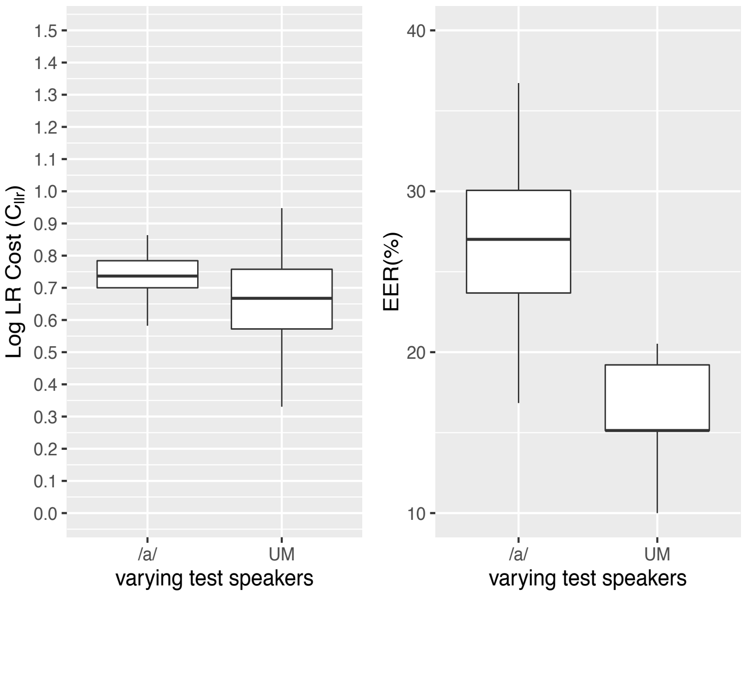

Figure 5: C_{llr} and EERs for /a/ and UM varying test speakers in each replication.

A somewhat different pattern is found for EER (right panel in Figure 5). Wider ranges of EER are found for /a/ than for UM. The overall and interquartile range for /a/ are 19.9% and 6.4% respectively, while they are 10.5% and 4.1% for UM. Again, the overall speaker discriminatory power is better for UM than for /a/. However, taken together with the C_{llr} results, the EER results suggest that while UM produces fewer errors overall, and more stable performance, there is greater variability in the magnitude of the errors that it produces.

	C_{llr}		EER (%)	
	/a/	UM	/a/	*UM*
Minimum	0.58	0.33	16.8	10.0
1st quartile	0.70	0.57	23.7	15.1
Median	0.73	0.67	27.0	15.1
3rd quartile	0.78	0.76	30.1	19.2
Maximum	0.86	0.94	36.7	20.5
Interquartile range	0.08	0.19	6.4	4.1
Overall range	0.28	0.61	19.9	10.5

Table 2: Minimum, maximum, median, mean, first and third quartiles of C_{llr}s and EERs of /a/ and *um* in experiment 2.

4.3 Experiment 3: Varying reference speakers

In experiment 3, training and test speakers were fixed across the 100 replications. A different set of reference speakers was sampled in each replication. The results are summarised in Table 3. The left panel of Figure 6 shows the C_{llr} boxplots for /a/ and UM. In both cases, the range of values is extremely small. /a/ yielded an overall Cllr range of between 0.66 and 0.77, while the interquartile range was 0.03. UM yielded an overall range of 0.14 and an interquartile range of 0.03. The right panel of Figure 6 shows the EERs. EER ranges from 23.2% to 30.5% for /a/ and from 10.0% to 15.1% for UM. The interquartile range for /a/ (1.7%) is lower than that for UM (4.1%), while the overall range of /a/ (7.6%) is higher than that of UM (5.1%). As in the previous experiments, general speaker discriminatory power is much better for UM than for /a/, with the EERs for UM lower than those of /a/ in each of the 100 replications.

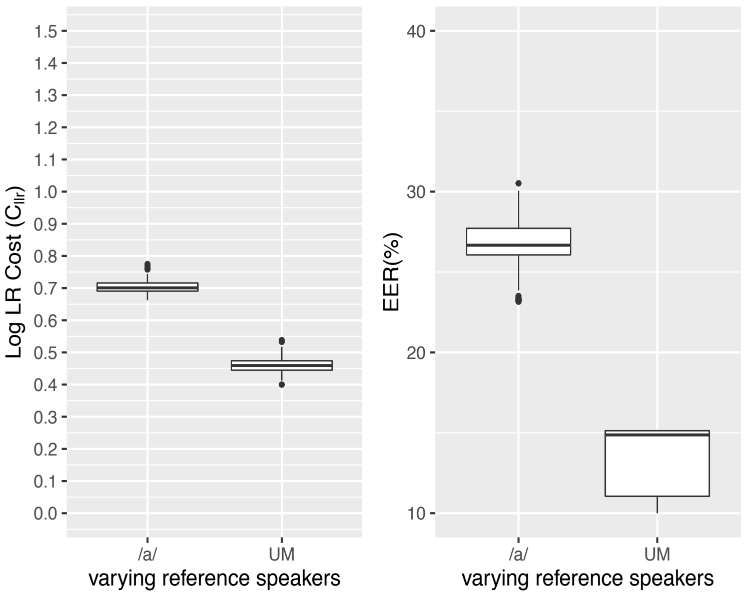

Figure 6: C_{llr}s and EERs for /a/ and UM varying reference speakers in each replication.

It is worth noting that the EER boxplot of UM has a very short lower whisker and no upper whisker. This is because the test speakers were fixed and, as such, there are only a discrete number of EERs possible given the number of speakers used. The smaller the number of speakers used, the fewer the number of possible EERs. Most of the EERs for UM are gathered around 14.9%.

	C_{llr}		EER (%)	
	/a/	UM	/a/	UM
Minimum	0.66	0.40	23.2	10.0
1st quartile	0.69	0.44	26.1	11.1
Median	0.70	0.46	26.7	14.9
3rd quartile	0.72	0.47	27.7	15.1
Maximum	0.77	0.54	30.5	15.1
Interquartile range	0.03	0.03	1.7	4.1
Overall range	0.11	0.14	7.6	5.1

Table 3: Minimum, maximum, median, mean, first and third quartiles, interquartile ranges and overall ranges of C_{llr}s and EERs for /a/ and UM in experiment 3.

Experiment 3 shows that using different reference speakers produces fairly stable system performance compared with Experiments 1 and 2. UM and /a/ yielded a similar system stability in terms of both C_{llr} and EER. Comparing with experiment 2, the system stability is less sensitive to different make-ups of reference speakers as long as they come from a matched dialectic group.

4.4 Experiment 4: Vary training speakers

In experiment 4, test and reference speakers were fixed throughout the 100 replications. Speaker randomisation was carried out by assigning a different set of speakers into the training data in each replication. Figure 7 shows the C_{llr}s for /a/ and UM across the replications, and the distributions are summarised in Table 4. The EER was not reported in detail in experiment 4, because the calibration coefficients derived from the training data only affect the C_{llr}. Thus, the

EER would be the same across all replications where the test and reference speakers are fixed. Both variables generally produced a very narrow ranges of C_{llr} values when varying the training speakers; narrower than ranges reported in experiments 1, 2 and 3. The overall and interquartile ranges of C_{llr}s are 0.12 and 0.01 for /a/. For UM, the interquartile range was 0.05. The overall range for UM, however, was 0.4 due to an outlying replication producing a C_{llr} 0f 0.89.

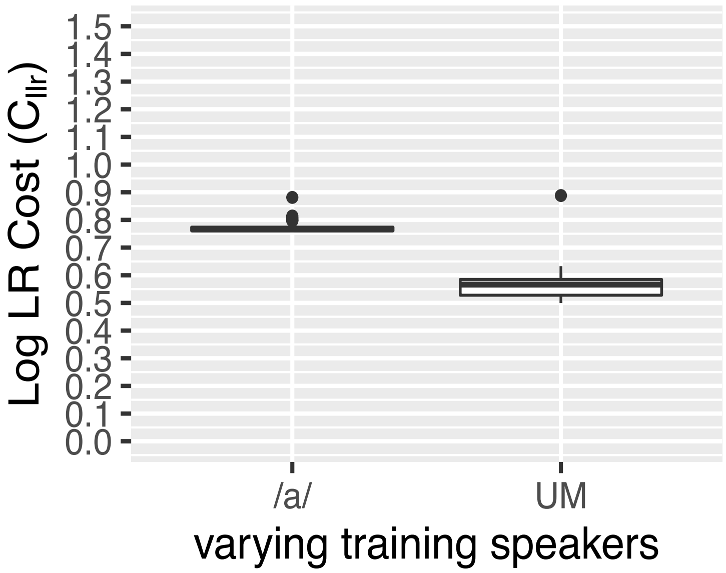

Figure 7: C_{llr}s of /a/ and `um` by varying training speakers in each replication.

	C_{llr}	
	/a/	UM
Minimum	0.76	0.49
1st quartile	0.76	0.53
Median	0.77	0.57
3rd quartile	0.77	0.58
Maximum	0.88	0.89
Interquartile range	0.01	0.05
Overall range	0.12	0.40

Table 4: Minimum, maximum, median, mean, first and third quartiles, interquartile ranges and overall ranges of C_{llr}s of /a/ and UM in experiment 4.

Experiment 4 shows that varying training speakers has a very limited effect on system stability. /a/ produced a marginally more stable system performance than UM in terms of C_{llr}. This is possibly due to the fact that there were only 20 training speakers used for UM, compared with the 30 training speakers used for Cantonese /a/.

5 Discussion

In this section, the results from the four experiments are discussed. Figure 8 shows the C_{llr} (upper panels) and EER (lower panels) values across the 100 replications for Cantonese /a/ and SSBE UM from the four experiments. However, the discussion below focuses on C_{llr} as it is the most widely used metric to evaluate system performance and stability and provides the most coherent story across experiments.

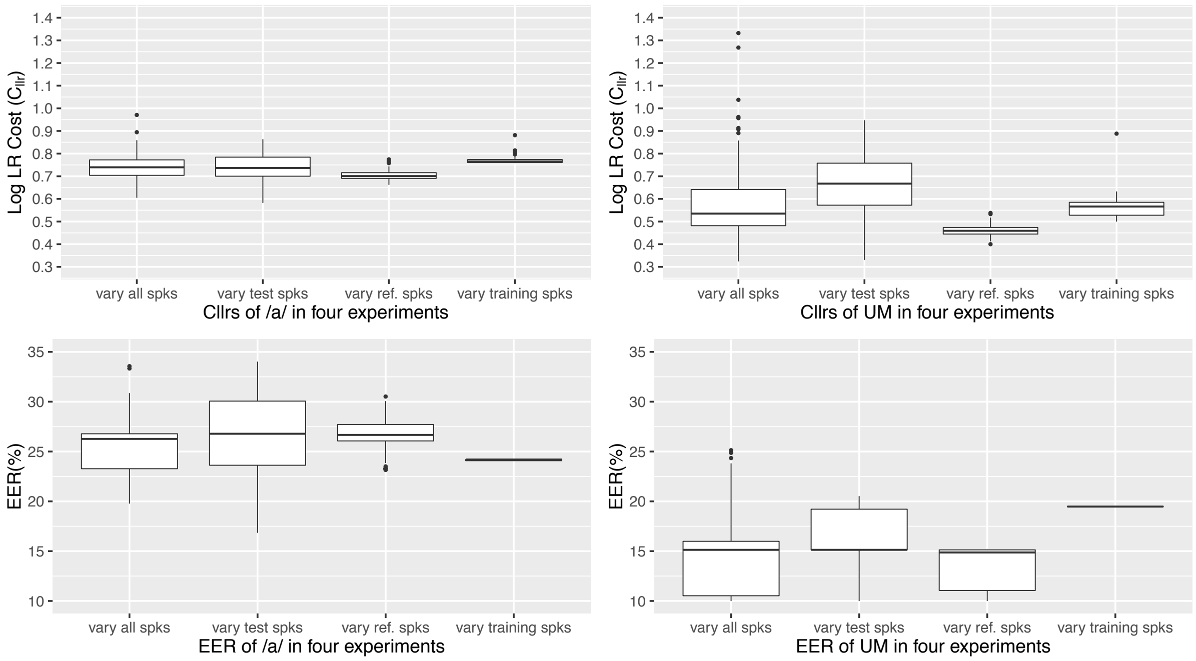

Figure 8: C_{llr} (top) and EER (bottom) values across the four experiments for /a/ (left) and UM (right).

General findings

Considerable variation was found across replications when varying the speakers in all three sets, with C_{llr} values ranging from 0.6 to 0.97 for Cantonese /a/ and from 0.32 to 1.33 for SSBE UM. The variability in performance, especially for UM, is as wide as the variability one would expect to see between variables and between populations; the difference between what would be considered very good performance versus very bad performance. Therefore, these results show that caution should be exercised when judging the speaker discriminatory power of a variable based on a single configuration of speakers in the training, test and reference sets. This is especially true since many LR-based studies use the same number of (or fewer) speakers than used here. Using different, but still representative, speakers could affect system performance substantially, depending on who exactly those speakers are. However, further examination of the results from the final three experiments in this study show that variability in system performance is almost exclusively due to the effects of varying the test speakers. For both variables, the range of C_{llr} values when varying only the test speakers was almost as wide as that when varying speakers in all three sets. By comparison, the variability in C_{llr} as a function of the make-up of the reference and training sets was extremely small. Thus, as long as the sets are of a sufficient size (in these experiments, over 20 speakers), the specific speakers used for the training and reference data have little effect on the overall performance of the system.

There were some differences in the results for the two variables examined here. Cantonese /a/ was less sensitive to speaker sampling than SSBE UM, reflected in a narrower range of C_{llr} values across the four experiments. One reason for this may be the inherent speaker-discriminatory power of /a/ compared with UM. The median C_{llr} value for UM in experiment 1 was 0.54, compared with 0.74 for Cantonese /a/. The lower end of the distribution of C_{llr} values for UM also shows that it has the potential to produce very good performance with certain configurations of speakers. This shows that inherently better speaker discriminants will produce more variable system performance because they have the potential to produce a wider range of results depending on which speakers are being used (principally, in the test set). However, poorer speaker discriminants, such as /a/, will produce poor system performance irrespective of the speakers used. This relationship between speaker discriminatory power and sensitivity to speaker sampling replicates findings reported in Wang et al. (2019) based on simulated data. A further contributing factor may also be that more speakers were available for the analysis of /a/ (30 speakers per set) compared with UM (20 speakers per set). It is likely that system performance is more stable when the number of speakers is higher.

Implications

The results presented here have important implications for issues of uncertainty, decision-making, and subjectivity and objectivity in data-driven forensic comparison. Specifically, they provide new insights on how best to go about testing and validating FVC systems in casework in order for the results to be useful to both the expert, and more importantly, to the trier-of-fact.

The general finding relating to the stability of system validity when using different sets of training and reference speakers is extremely positive for casework. The training and reference sets are the core elements of any system; the test data are simply intended to be 'unseen' representative speakers to assess how well the system performs, and are not part of the system *per se*. The fact that performance is so insensitive to speaker variability in the training and reference sets means that the expert can be relatively confident about the transferability of the system as long as the speakers used are broadly representative of the relevant population, i.e. the uncertainty as a function of the make-up of the training and reference sets is low. The lack of variability in system performance when varying the training and reference sets is all the more impressive given that we are working with relatively small, manageable numbers of speakers (as low as 20 per set).

However, the key source of uncertainty in system performance derives from the make-up of the test set. Therefore, it is essential that the expert carefully considers the speakers that are used for testing, since this can have a substantial effect not only on the system that an expert decides to use in a case, but also may over- or under-estimate the true validity to the court, potentially leading to incorrect decisions being made by the trier-of-fact. One way in which to deal with this issue is to use data that are more representative of the 'type' of voices in the case – the issue here is not one of finding the best configuration of test speakers to produce the optimal performance, but rather to find a set that produces a validity measure that is representative for the case. It has been argued that systems should be evaluated using recordings that reflect the conditions of the case at trial (Enzinger and Morrison 2017; Enzinger, Morrison and Ochoa 2016). In terms of speaker characteristics, this is taken to mean that the speakers used are representative of the relevant population, often defined broadly by sex and language (Rose 2004). Clearly, based on the variability reported in our studies, this is insufficient, especially where the variable(s) can potentially provide good speaker discrimination and/or the number of speakers is small. Identifying a subset of a database of speakers who are in some way 'more similar' to the offender (akin to the suggestion in

Morrison, Ochoa and Thiruvaran 2012, and procedures in some automatic systems) is likely to produce more representative results. However, this involves making pragmatic but subjective decisions: either based on narrower demographic properties of the offender (although, of course, we can't know these properties for certain, since the identity of the offender is the very question at stake) or using some measure to define speaker similarity. This is not necessarily problematic. As highlighted by the court in R v T [2010] "the probability that is quoted (by the expert as a conclusion in a case) … will inevitably be a personal probability and the extent to which the data influence that probability will depend on expert judgement" (at para 80).

Having tailored test data is, in our view, the preferred approach to reducing the uncertainty in the performance of a system. This allows the expert to have a better sense of how the system will perform in the specific case. However, as highlighted in Hughes and Foulkes (2015), there will always be some mismatch between the data used for building and evaluating a system and the case data. It is likely, therefore, to be fruitful to examine ways to further reduce uncertainty by incorporating it into the LR computation itself; for an example based on sample size see Morrison and Poh (2018). Alternatively, we consider it a minimal requirement that both researchers and experts undertake speaker sampling of the kind described in this study in order to understand the potential range within which a system performs. This may not provide case-specific information, but will provide insights into how certain we can be about the performance of a system in general. For instance, the range of values produced for UM in this study means that we would need to be extremely cautious about making generalisations about speaker discriminatory power or the usefulness of such a system in casework.

6 Conclusion

The current study has explored the effect of speaker sampling in LR-based FVC through four experiments. Experiment 1 explored the stability and reliability of both the system and segmental variables, which shows that the same segmental variable might give very different performance just by rearranging speakers in the training, test and reference data. Experiments 2 to 4 examined the variability in system performance when separately varying the training, test and reference speakers. Results showed that the key source of variability in system performance is the make-up of the test set. This is positive for casework since it suggests that the make-up of the training and reference sets, the key elements of a system, have little effect on system performance.

However, as discussed by Curran (2016) and others, consideration should also be given to dealing with uncertainty in system performance. Here we see the importance of uncertainty specifically in the test set, which can be established either by using a more tailored subset of test speakers or, minimally, reporting the range of values produced through speaker sampling of the sort described here. In this way the analyst can provide an estimate of the range of validity values the system can produce, and thus provide a means to record the precision or uncertainty in the LR calculations. Providing such information is critical for the trier-of-fact to evaluate the evidence provided by the expert.

7 Acknowledgements

We would like to thank the editor and an anonymous reviewer for their constructive comments We are also grateful to Justin Lo for the R scripts which facilitated data analysis and the first author takes full responsibility for any potential mistakes that might occur.

Reference

Aitken, C. G., and 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.

Andrus, Tony, et al. (2016) IARPA Babel Cantonese Language Pack IARPA-babel101b-v0.4c LDC2016S02. Web Download. Philadelphia: Linguistic Data Consortium.

Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.36.

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

Chen, A., and Rose, P (2012) Likelihood ratio-based forensic voice comparison with the Cantonese triphthong/iau. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*: 197-200.

Curran, J. M. (2016) Admitting to uncertainty in the LR. *Science and Justice,* 56: 380-382.

Enzinger, E. and Morrison, G.S. (2012) The importance of using between- session test data in evaluating the performance of forensic-voice- comparison systems. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*: 137–140.

Enzinger, E., Morrison, G.S., and Ochoa, F. (2016) A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science and Justice*, 56: 42-57.

Enzinger, E. and Morrison, G.S. (2017) Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International*, 277: 30-40.

Grigoras, C., Smith, J., Morrison, G. and Enzinger, E. (2013) Forensic audio analysis – Review: 2010-2013. *Proceedings of the 17th International Science Managers' Symposium*: 612– 637.

Home Office. (2003). Criminal Justice Act (Chapter 44). Her Majesty's Stationery Office.

Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough?. *Speech Communication,* 94: 15-29.

Hughes, V. and Foulkes, P. (2017) What is the relevant population? Considerations for the computation of likelihood ratios in forensic voice comparison. *Proceedings of Interspeech*: 3772-3776.

Hughes, V. and Foulkes, P. (2015) The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication,* 66: 218-230.

Hughes, V., Wood, S. and Foulkes, P. (2016) Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law,* 23(1): 99-132.

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

Kinoshita, Y. and Ishihara, S. (2014) Background population: how does it affect LR-based forensic voice comparison? *International Journal of Speech, Language and the Law,* 21(2): 191-224.

Kwok, H. (1984) *Sentence particles in Cantonese*. Centre of Asian Studies: University of Hong Kong.

Law, A. (2002) Cantonese sentence-final particles and the CP domain. *UCL working papers in linguistics*, 14: 375-398.

Lennes, M. (2003a) 'Save_intervals_to_wav_sound_files.praat'. Retrieved on 21August 2018 from https://github.com/FieldDB/Praat-Scripts

Lennes, M. (2003b) 'Collect_formant_data_from_files.praat' Retrieved on 21August 2018 from https://github.com/FieldDB/Praat-Scripts

Leung, W. M. (2009) A study of the Cantonese hearsay particle wo from a tonal perspective. *International Journal of Linguistics,* 1(1): 1-14.

Lindblom, B. (1963) Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America,* 35(11): 1773–1781.

Liu, X. M. (2006). *刑事侦查程序理论与改革研究 [Criminal investigation theory and reform].* China Legal Publishing House.

Lo, J. (2018). FVClrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison [unpublished R package]. https://github.com/justinjhlo/fvclrr

Matlab implementation of Aitken and Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation. MORRISON, G.S. (2007). Retrieved on 20 July 2018 from Geoff-morrison.net/#MVKD

McDougall, K. (2004) Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law*, 11(1): 103-130.

McDougall, K. (2006) Dynamic features of speech and the characterisation of speakers: towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law*, 13(1): 89-126.

Morrison, G. S. (2008) Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /ai/. *International Journal of Speech, Language and the Law*, 15(2): 249-266.

Morrison, G. S. (2009) Likelihood-ratio voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America*, 125(4): 2387-2397.

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

Morrison, G. S. (2011) 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(2): 242-256.

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. (2016) Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Science and Justice*, 56(5): 371-373.

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

Morrison, G. S. and Poh, N. (2018) Avoiding overstating the strength of forensic evidence: shrunk likelihood ratios/Bayes factors. *Science and Justice*, 58: 200-218.

Nolan, F., McDougall, K., de Jong, G. and Hudson, T. (2009) The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law,* 16(1): 31-57.

R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Robertson, B. and G. A. Vignaux. (1995b) *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Oxford: Oxford University Press.

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

Rose, P., and Morrison, G. (2009) A response to the UK position statement on forensic speaker comparison. *The International Journal of Speech, Language and the Law*, 16(1): 139.

Rose, P., and Wang, X. (2016) Cantonese forensic voice comparison with higher-level features: likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone. *Proceedings of Odyssey:* 326-333.

Shriberg, E. (2001) To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31: 153-169.

Sybesma, R., and Li, B. (2007) The dissection and structural mapping of Cantonese sentence final particles. *Lingua, 117*(10): 1739-1783.

Tschäpe, N., Trouvain, J., Bauer, D., and Jessen, M. (2005) Idiosyncratic patterns of filled pauses. Proceedings of *14th Annual Conference of the International Association for Forensic Phonetics and Acoustics,* Marrakesh, Morocco.

Wang, B., Hughes, V. and Foulkes, P. (2019) Effect of score sampling on system stability in likelihood ratio based forensic voice comparison. In *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne, Australia.

Zhang, C., Morrison, G. S. and Thiruvaran, T. (2011) Forensic voice comparison using Chinese/iau/. In *Proceedings of the 17th International Congress of Phonetic Sciences*, 17: 21.