---

**Article:**
Wang, Xiao, Hughes, Vincent orcid.org/0000-0002-4660-979X and Foulkes, Paul orcid.org/0000-0001-9481-1004 (2022) The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison. Speech Communication. ISSN: 0167-6393

---

Title: The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison

Bruce Xiao Wang*, Vincent Hughes, Paul Foulkes

*Department of Language and Linguistic Science, University of York, Heslington, York YO10 5DD, United Kingdom*

*Corresponding author.
E-mail address: xw961@york.ac.uk (Bruce Xiao Wang)

Abstract

The likelihood ratio (LR) framework has been widely adopted in voice (and other forensic) evidence evaluation. However, in developing any forensic comparison system, it is necessary to make subjective and pragmatic decisions, which in turn may affect the results that system produces. One such decision relates to not only the size of the samples used for training and testing the system, but also which specific individuals are used in the samples. The current study explores the relationship between sampling variability (i.e. the choice of speakers used for training and testing systems, rather than sample size) and the choice of linguistic features used. The first three formants and f0 from the vocalic portion of the filled pause *um* were used as input, as well as both vowel and nasal durations. 25 speakers were used in test, training and reference sets respectively. Experiments were carried out using all 31 logically possible combinations of features, and replicated 100 times using different configurations of 25 training and reference speakers. The results show that a) overall, $C_{llr}$ mean reduces with more features involved and no clear pattern is observed in $C_{llr}$ range; meanwhile, considerable fluctuation is observed within individual speakers; b) while the majority of speakers yield stronger mean LLRs in systems with three or more features, a few speakers can be well-separated using one or two features; c) sampling variability in the training and reference speakers has limited effect on individual test speakers' LR outputs in same-speaker (SS) comparisons, but a marked effect on different-speaker (DS) LRs.

Key words: forensic phonetics, likelihood ratio, sampling variability, individual behaviour

# 1 Introduction

During the past few decades, the likelihood ratio (LR) framework has been widely adopted in forensic voice comparison (FVC) and other forensic sciences for evidence evaluation (Champod & Meuwly, 2000; Morrison, 2009b; Nolan, 2001; Rose & Morrison, 2009). The numerical LR framework relies on empirical data to estimate the strength of evidence and evaluate the validity and reliability of the system used to perform the comparison. Two crucial issues lie at the heart of such work: whether the LR is assessing what the analyst wants it to (validity), and whether one would get the same LR value if the comparison is repeated (reliability). Numerous empirical studies have explored the validity and reliability of LR output as a function of sample size (e.g. Hughes, 2017; Ishihara & Kinoshita, 2008; Kinoshita & Ishihara, 2015), sampling variability (e.g. Ali et al., 2015; Wang et al., 2019) and the choice of linguistic features (e.g. Hughes et al., 2016; Morrison, 2009a; Zhang et al., 2011). However, the majority of this research places a strong emphasis on assessing *overall* performance. By contrast, few have explored the behaviour of individual speakers within these systems (but see e.g. Lo, 2021) and the interaction between sample size, sampling variability and choice of features. Given what really matters in a real case is the specific voice of the speaker(s) under analysis, this raises issues about whether *overall* performance provides adequate insight into the validity and reliability of an LR system in a specific case, because we are often dealing with small sample sizes in the real world. In computing an LR, potential variability could be caused by a range of factors (see Morrison, 2016 for details). The current study explores the relationship between overall performance and individual speakers' behaviour in relation to two sources of variability: sampling variability in the relevant population, and variability in the choice of linguistic features. The following sections give some background information on LR-based FVC and previous studies in system testing. Detailed experimental methods are then

presented in section 2, followed by results (section 3), discussion (section 5) and conclusion (section 6).

### 1.1 Likelihood ratio based forensic voice comparison

In forensic voice comparison (FVC), a typical scenario is to compare recordings, one of an unknown offender, and the other of a known suspect (in the UK the suspect recording is from a police interview). The role of the expert is to evaluate the suspect and offender speech samples to assist the trier-of-fact in determining the likelihood that the two speech samples came from the same person or different people. The linguistic-phonetic method is widely employed in casework around the world, where segmental (e.g. vowels, consonants), suprasegmental (e.g. f0, intensity, speech rate), and non-linguistic features, such as hesitation markers, are analysed (Gold & French, 2011, 2019). Applying the LR framework to FVC means that experts need to estimate the strength of evidence under the two competing propositions of the prosecution and defence (Robertson et al., 2016), i.e. what is the probability of observing the speech evidence had it come from the same speaker or different speakers ? The LR equation is expressed in [1]:

$$LR = \frac{p(E|H_p,I)}{p(E|H_d,I)} \quad [1]$$

where $p(E|H_p)$ indicates the probability of the evidence given the suspect and offender samples come from the same person; $p(E|H_d)$ represents the probability of the evidence given the suspect and offender samples come from different people; $I$ stands for background information about the case. Essentially, the numerator of the LR is an estimation of the similarity between the suspect and offender speech patterns, while the denominator is an estimation of their typicality in relevant population, i.e. what is the probability of observing the

offender's speech pattern had it not come from the suspect but some other randomly selected member of the relevant population?

In order to generate a LR (be that numerical or verbal; or indeed any form of conclusion in a FVC case), the expert employs a *system*. This is defined broadly as the particular courses of action that are used to compare the suspect and offender samples (Morrison, 2013), e.g. data used to represent the relevant population, variables chosen for analysis, methods of analysing those variables, statistical models used for score generation, and calibration methods used for score-to-LR conversion. For an end user (e.g. court) to be able to interpret the conclusion provided by the expert appropriately, it is essential to understand the validity of the system used to generate that conclusion. The key question, therefore, is: to what extent does the system perform the task that it is designed to do? There is now growing pressure from within the forensic sciences and from external government and regulatory bodies for experts to validate their systems empirically using data where the ground truth is known, and to present the results of validation tests to the end user. Within data-driven LR-based FVC (and across other disciplines) there is now a widely accepted framework for validating systems (Morrison et al., 2021). This involves two stages and three separate datasets, namely, test, training and reference datasets. In stage one (the *feature-to-score* stage), pairs of same-speaker (SS) and different-speaker (DS) recordings taken from the test and training datasets are compared to produce training and test scores which indicate the similarity between the SS and DS samples, and assessing typicality with respect to a reference population. In stage two (the *score-to-LR* stage), the training scores are used to generate calibration coefficients which are then applied to the test scores to convert them to interpretable LRs. System validity and reliability metrics are generated from the calibrated LRs for the test set.

### 1.1.1 Feature selection and overall performance

Numerous previous studies have looked at the performance of FVC systems using different linguistic features as input, for example, Morrison (2009a) using the Australian diphthong /aɪ/; Zhang et al. (2011b) using the Mandarin triphthong /iau/, and Chen & Rose (2012) using the Cantonese triphthong /iau/. Polynomial curves were fitted to segmental features, and the coefficients of polynomial curves were used as the input for LR computation. Moreover, some studies also investigated the speaker-discriminatory power using consonants, for example Rose (2003) with Japanese nasals, Rose (2011) with Japanese fricatives, and Kavanagh (2012) with English laterals. Other studies have looked into the speaker-discriminatory power using suprasegmental variables, e.g. Cantonese tonal f0 (Rose & Wang, 2016), long-term f0 distribution (Kinoshita et al., 2008), voice quality (Hughes et al., 2019), and articulation rate (Lennon et al., 2019). In general, there is the tendency in the literature that higher formants outperform lower formants.

### 1.1.2 Sampling variability, sample size and overall performance

Sample size is one of the potential sources of variability in LR output. In principle, the larger the sample, the more precise the output. However, in linguistic casework analysts often deal with small sample sizes. The question for real casework then is, how do systems perform with limited data? Previous studies have explored overall performance as a function of the number of test, training and reference speakers used. For example, Ishihara & Kinoshita (2008) and Kinoshita & Ishihara (2014) used long-term distribution of f0 data (i.e. the overall shape of long-term f0 distribution) from 241 Japanese speakers. The $C_{llr}$ and EER reduced from ca. 0.7

to 0.4 and 17.2% to 8% respectively as the number of reference speakers increased from 10 to 120. Due to the limited number of recordings per speaker, the credible interval (a Bayesian measure of reliability; Morrison, 2016) was calculated for DS comparisons only. Results showed that there was not much improvement in system precision once the number of the reference speakers reached 30. Similarly, Hughes (2017) used simulated data to investigate overall performance as a function of number of test, training and reference speakers. Midpoint F1, F2 and F3 values were simulated using the acoustic data measured for the filled pause *um* from 86 male speakers of Standard Southern British English. The system was tested by varying the number of speakers from 2 to 100 for test, training and reference sets respectively. The results from Hughes (2017) showed that system validity was more affected by the number of test and training speakers and less affected by the number of reference speakers. A relatively stable system can be achieved with large amounts of training and test speakers (e.g., 30 to 40 speakers per set) and a moderate number of reference speakers (e.g., 15 speakers).

Wang et al. (2019) showed that overall performance is not only affected by the number of speakers used in test, training and reference sets, but also which speakers are used in each set. They explored overall performance as a function of the configurations of test, training and reference speakers. It was predicted that stable systems would yield consistent results regardless of the different configurations of speakers in each data set. 155 Cantonese speakers and 73 SSBE speakers were used. The Cantonese sentence final particle (SFP) /a/ and English FP *um* were used as the linguistic variables, and fixed sets of acoustic features (F1 and F2 for /a/ and F1, F2 and F3 for *um*) were used as input for computing numerical LRs. 30 speakers were used for each of the test, training and reference sets for Cantonese data, while 20 speakers were used for each of the test, training and reference sets for English data. Four experiments were run, and each replicated 100 times. In each replication the choice of speakers in the data

sets was randomly varied. In one experiment, test, training and reference speakers were all varied across replications. This mimics research practice in LR-based FVC where normally only one configuration of test, training and reference speakers is used. In the other three experiments only one of the test, training or reference groups was varied across replications. In varying only one of the data sets, the sensitivity of the overall performance was assessed in relation to different configurations of speakers in each data sets respectively. Results show that both Cantonese SFP /a/ and English FP *um* yielded the lowest system stability when varying speakers in all three datasets, with the $C_{llr}$ values ranging between 0.60 to 0.97 for /a/ and 0.32 to 1.33 for *um*. However, the variability was primariy caused by different configurations of test speakers. The $C_{llr}$ ranged from 0.58 to 0.86 for /a/ and 0.33 to 0.94 for *um* when only test speakers were varied, while varying only training or only reference speakers had a limited effect on system stability. Given 1 is the logical threshold for $C_{llr}$ (Morrison et al., 2021), i.e. systems with $C_{llr}$ higher than 1 do not give much useful information, the overall performance shows a markedly wide range by simply randomly select different test speakers from the relevant population.

## 1.2 The current study

Most previous studies have placed a strong emphasis on generic system testing (i.e. testing different linguistic/signal processing features for lower $C_{llr}$/EER). The current study explores the relationship between overall performance and individual speaker behaviour under two different conditions. First, we explore how overall performance and the LRs for individual speakers are affected when different features and sets of features are used; in this regard we compare single-feature systems (using individual vowel formants F1, F2 or F3) and various multi-feature systems (using different combinations of formants, e.g. F1 and F2 or F2 and F3).

This is because, in real cases, the speech data available for analysis varies (Gold & French, 2011) and not all features are reliable in different conditions, e.g. the estimates of F1 are artificially raised by telephone transmission by an average of 14% for landlines (Künzel, 2001) and 29% for mobiles (Byrne & Foulkes, 2004). Second, we explore how system and individual speakers' results are affected by the random effects of inter- and intra-speaker variation within the relevant population. Throughout our experiments, the test speakers are held constant (i.e. the same test speakers are used with different configurations of training and reference speakers) to allow analysis of results across the same set of comparisons (see 2.2 for detailed experimental procedure).

## 2 Method

### 2.1 Speech materials and feature extraction

The data used in the current study is an extension of that used by Hughes et al. (2016), containing the FP *um* from 90 SSBE speakers (drawn from the DyViS corpus; Nolan et al., 2009). All speakers were aged between 18 and 25. The FP *um* was extracted from spontaneous non-contemporary speech in two tasks, a mock police interview (Task 1) and a telephone conversation with an 'accomplice' (Task 2). All tasks were recorded in high-quality studio conditions (44.1 kHz sampling rate, 16-bit depth).

For each token, the data consists of nine raw Hz measurements for each formant and f0 taken at +10% steps across the duration of the vowel, as well as nasal and vocalic duration (see Hughes et al. 2016 for detailed data extraction procedure). There are on average 35 FP *um* tokens per speaker per task. The first three formants and f0 of each token were fitted using

quadratic polynomial curves, as no single formant or f0 was expected or observed to display a more complex shape. Figure 1 shows an example of quadratic fitting to five tokens from speaker #114 in Task 1 (we refer throughout the paper to DyViS speaker codes in the form #114). The quadratic coefficients of the first three formants and f0 as well as the vocalic and nasal durations were then used as the input features for score generation and LR computation.

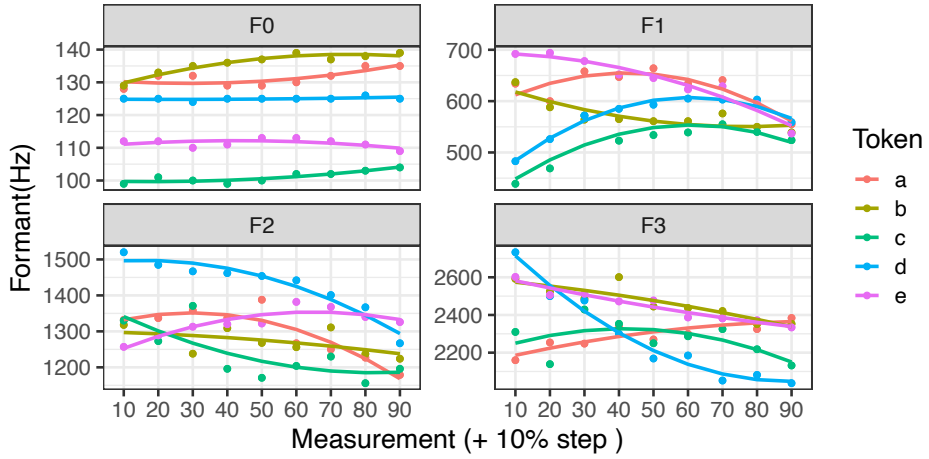


Figure 1. Quadratic curve fitting to F0, F1, F2 and F3 of five tokens from speaker #114 in DyViS Task 1.

## 2.2 Experiment procedure

25 speakers were randomly sampled to act as the test, training and reference data respectively. Task 1 and Task 2 were used as the nominal suspect and offender samples respectively. The SS and DS pairs of test and training data were compared using the MVKD formula (Aitken & Lucy, 2004) to produce test and training scores. The training scores were then used to train a logistic regression model (Brümmer et al., 2007; Morrison, 2011) which was then applied to the test scores to produce calibrated LRs. The implementation of the current experiment was carried out in R (R Core Team, 2020) using the `fvclrr` package (Lo, 2018). In order to

explore overall performance and individual speaker behaviour with respect to the choice of linguistic features and sampling variability in the relevant population, experiments were carried out using all 31 possible combinations of features. i.e. 5 single-feature systems, 10 two-feature systems, 10 three-feature systems, 5 four-feature systems and 1 five-feature system. Experiments were replicated 100 times using different configurations of training and reference speakers, but keeping the 25-speaker test set fixed. This enables us to assess the LRs for the same test speakers using different input features and different configurations of training and reference speakers. 50 out of the remaining 65 speakers were used in each replication to allow for different configurations of training and reference speakers. As such, overall sample size remained fixed throughout. The choice of using 25 speakers per set is driven by the size of the available database and realistic expectations for how much data would be available in casework. Indeed, in much FVC research, it is typical for sets of around 20 speakers to be used (e.g. Zhang et al., 2013). A schematic diagram of experimental procedure is given in Figure 2.

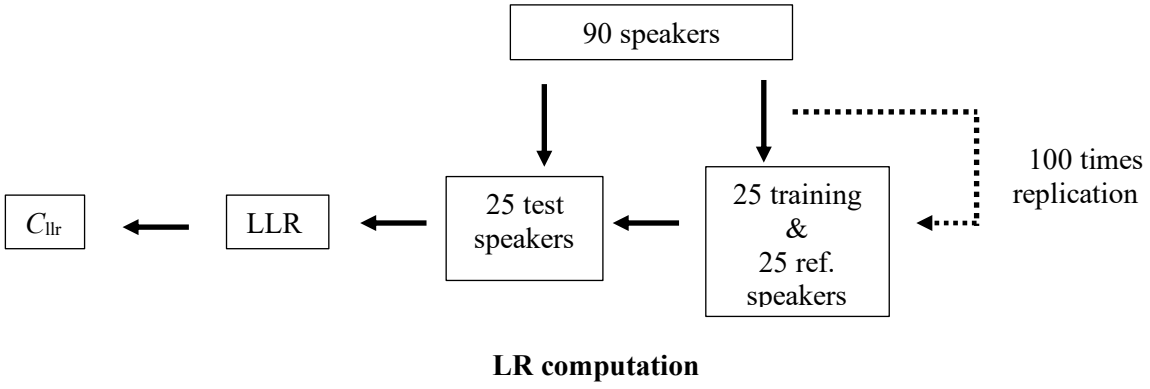


Figure 2. A schematic diagram of experiment procedure in LR computation.

It is acknowledged that the recordings used were high quality speech and the training and reference speakers across 100 replications were not independent of each other. This is likely to underestimate the variability in system output, compared with using low quality non-

contemporaneous recordings and truly independent samples of speakers. However, given the limitations on the availability of ideal data, it is likely that any type of replication study of this sort would use samples which are not entirely independent. Therefore, our results should be treated as 'base case scenario' for variability in system performance as a function of sampling. Even wider variability in system performance can be expected where poor quality recordings are used and independent samples are drawn from a much larger database.

## 2.3 Evaluation

Overall performance was evaluated using the log LR cost function ($C_{llr}$; Brümmer & du Preez, 2006), and the mean and range of $C_{llr}$ values were used to assess the system validity and stability respectively. A $C_{llr}$ threshold of 1 was used for determining whether the system is capturing any information (Morrison et al. 2021). To avoid confusion, the term *stability* is used, instead of *reliability*. Individual speaker behaviour was assessed using mean $Log_{10}$ LR (LLR) and root-mean-square-error (RMSE) values with reference to the LR verbal scale (Table 1; Champod & Evett, 2000). Although we acknowledge the drawbacks in using verbal scales, e.g. cliff edge effects, subjective interpretation (See Marquis et al., 2016; Morrison & Enzinger, 2016 for more discussion on verbal and numerical scales). RMSE error is used to measure the overall variability in LLRs across replications. The RMSE equation is expressed in equation [2], where $n$ is the total number of SS or DS comparisons, $x_i$ is the SS/DS LLR of individual speakers in each comparison, $y_i$ is the mean SS/DS LLR of individual speakers across 100 replications. The RMSE values of SS and DS comparisons were calculated for each test speaker in each system measuring how far each comparison LLR is from the mean for that speaker. A stable system and speaker would have low $C_{llr}$ range and low RMSE value respectively, which

indicates that the system and speaker are less sensitive to sampling variability in training and reference speakers.

| LLR | Verbal Expression |
|---|---|
| ±4: ±5 | Very strong support |
| ±3: ±4 | Strong support |
| ±2: ±3 | Moderately strong support |
| ±1: ±2 | Moderate support |
| 0: ±1 | Limited support |

Table 1. LLR verbal scale

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i - y_i)^2} \quad [2]$$

## 3 Results

### 3.1 Overall performance

The boxplots in Figure 4 show the distributions of $C_{llr}$ values across the 100 replications for each of the 31 systems. The x-axis displays the different systems, based on different combinations of input features. For example, 'F0' refers to the system where only f0 was used as input, while 'F01' indicates the combination of f0 and F1. The y-axis shows the $C_{llr}$ values. The top panel shows the $C_{llr}$ range of systems with one and two features, and the bottom panel shows the $C_{llr}$ range of systems with three, four and five features respectively.

For single feature systems, F2 yields much the lowest $C_{llr}$ mean (0.39) and range (0.04). The other four single feature systems yield similar overall performance, with the mean $C_{llr}$ around 0.7. F1 yields the least stable overall performance with an overall $C_{llr}$ range of 0.47, while F0, F3, and DUR (both vowel and nasal duration) systems yield a $C_{llr}$ range between 0.12 and 0.18.

Comparison across the single feature systems shows that F2 is the least sensitive to different configurations of training and reference speakers.

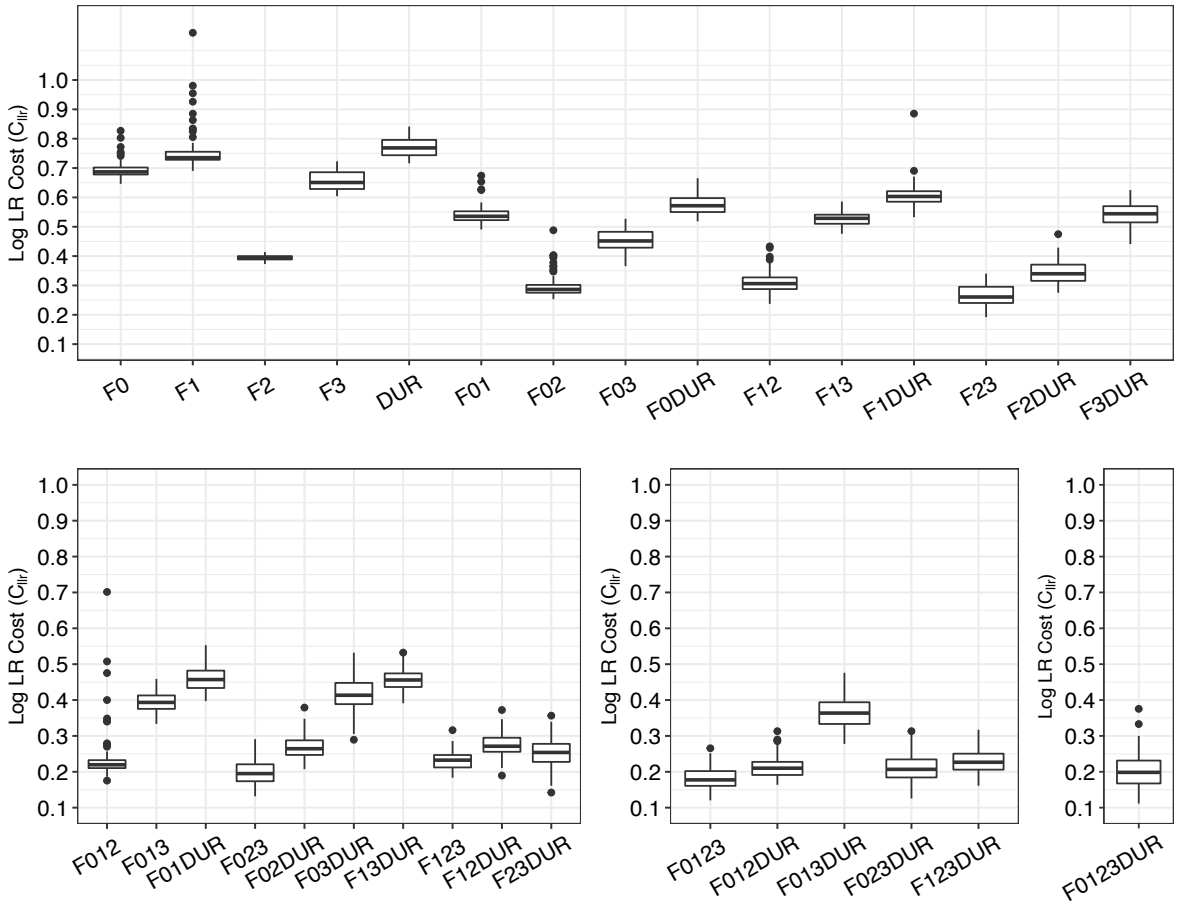


Figure 4. $C_{llr}$ variation across 31 systems (Top panel: systems with one or two features; bottom left: systems with three features; bottom middle: systems with four features; bottom right: system with five features).

The $C_{llr}$ patterns are more variable across systems with two features. The F23 system yields the lowest mean $C_{llr}$ (0.27), while the F0DUR system gives the highest mean $C_{llr}$ (0.64). In terms of system stability, the F13 system yields the lowest $C_{llr}$ range (0.11), and the F1DUR system yields the highest (0.35). A consistent pattern in the two-feature systems is that systems which include F2 outperform systems without F2 in terms of mean $C_{llr}$. In terms of the overall performance (i.e. lowest $C_{llr}$ mean and range), the F23 system is the best. It can be observed

that the F13 system has a lower $C_{llr}$ range than the F23 system; however, all of the $C_{llr}$ values in the F23 system are lower than those in the F13 system and the $C_{llr}$ range of the F23 system is only marginally higher than that of the F13 system. It is worth noting here that the effects seen here with regard to F1 cannot be explained by the telephone effect (Künzel, 2001) as all recordings are high-quality samples.

Among systems with three features, the F023 system yields the lowest mean $C_{llr}$ (0.20), while the F01DUR and F13DUR systems yield the highest (0.46). Similar to systems with single and two features, systems with F2 involved again outperform those without in terms of mean $C_{llr}$, e.g. the F012, F123, and F023 systems have lower mean $C_{llr}$ values than the F013 and F01DUR systems. In terms of $C_{llr}$ range, the F013 and F123 systems yield the lowest $C_{llr}$ range (0.13), while the F012 system yields the highest due to extreme outliers (0.53). The systems with the duration feature, i.e. F01DUR, F12DUR, F23DUR, F02DUR, F03DUR and F13DUR yield similar $C_{llr}$ range varying from 0.16 to 0.24. Overall, the F023 system has a marginally lower mean $C_{llr}$ and higher $C_{llr}$ range than the F123 system, and these two systems have similar overall performance and are less sensitive to different configurations of training and reference speakers than other systems.

For systems with four and five features, the F013DUR system gives the highest mean $C_{llr}$ (0.37), while other systems have similar mean $C_{llr}$ values. Combining all features does not improve overall performance, as the mean $C_{llr}$ (0.2) of the F0123DUR system is slightly higher than that of the F0123 (0.18) system, and the $C_{llr}$ range of the F0123DUR system (0.26) is higher than all other systems with four features.

Figure 5 shows the relationship between $C_{llr}$ mean (x-axis) and $C_{llr}$ range (y-axis) of the 31 systems. The general pattern shows that system validity and stability improve when more features are involved, i.e. systems with more features have a tendency of shifting to the bottom left corner. However, this is not always the case; for example, the system using F2 alone outperforms two-feature (e.g. F03, F13) and three-feature systems (e.g. F03DUR, F13DUR) in terms of both validity and stability; two-feature systems (F02, F12, F23) outperform three- and four-feature systems (e.g. F03DUR, F013DUR) in terms of validity (x-axis).

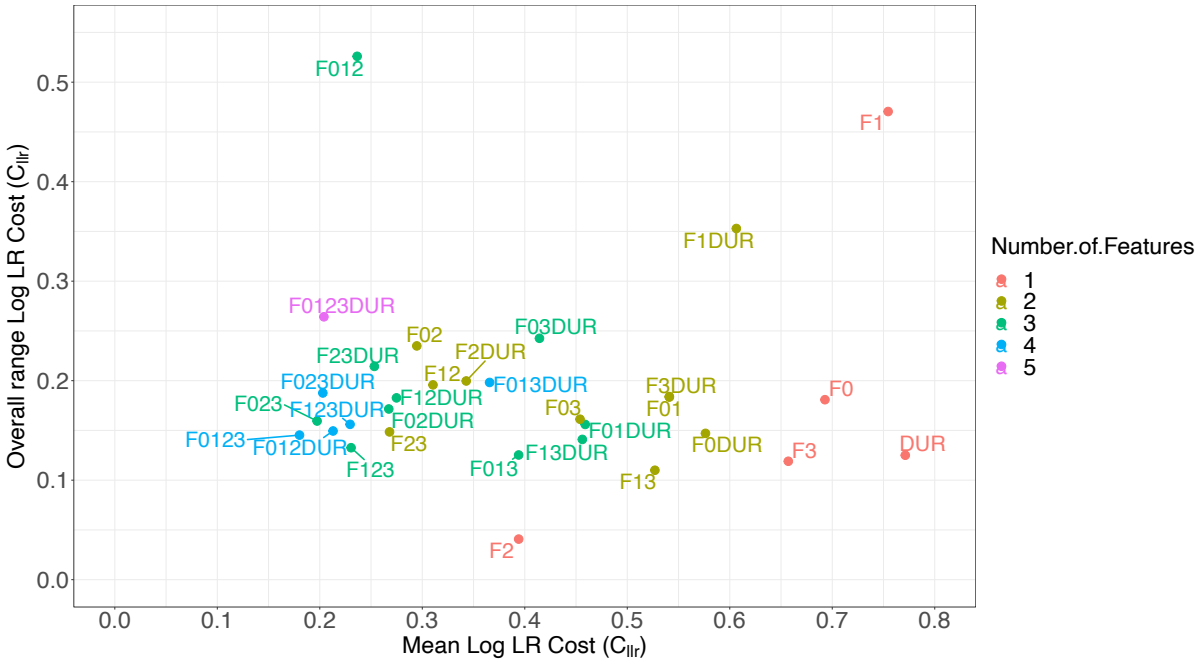


Figure 5. $C_{llr}$ mean and range across 31 systems (colour-coded according to the number of input features).

## 3.2 Individual behaviour

Since we also aim to explore how individual speakers behave when different features and configurations of training and reference speakers are used, we identified three specific speakers for further exploration whose results are representative of key patterns in the data.

Figure 6 shows the mean SS and DS LLRs for speakers #48, #51 and #53 across all 31 systems (see Appendix for all 25 speakers). The x- and y-axes display the SS and DS LLRs, and the legend indicates the number of features in the system. The vertical and horizontal lines indicate SS/DS LLR equal to 0 (the threshold between support for the prosecution and support for the defence on the LLR scale). Most speakers yielded a similar pattern to speaker #51, showing that speakers are more likely to yield more accurate performance (positive mean SS LLRs and negative mean DS LLRs) in systems with more features, i.e. the crosses, black circles and black triangles have a tendency to shift to the top right corner. The majority of speakers also produced consistent-with-fact results in systems with three or more features. However, exceptions can be found, as illustrated by speakers #48 and #53. Systems with four or five features do not produce stronger mean LLRs for these speakers. Instead, #48 and #53 can be well separated using a two-feature system (i.e. F02; indicated by arrows in Figure 6 lower panels). The maximum magnitude of LLRs exceeds 2.3 for SS comparisons, equivalent to *moderately strong support* for $H_p$, while the DS LLR is over -20, equivalent to *very strong support* for $H_d$. Moreover, speaker #48 produced contrary-to-fact LLRs in SS comparisons in the F013 (SS LLR = -0.38) and F13DUR (SS LLR = -0.02) systems and speaker #53 produced contrary-to-fact results in the F13DUR system (SS LLR = -0.42).

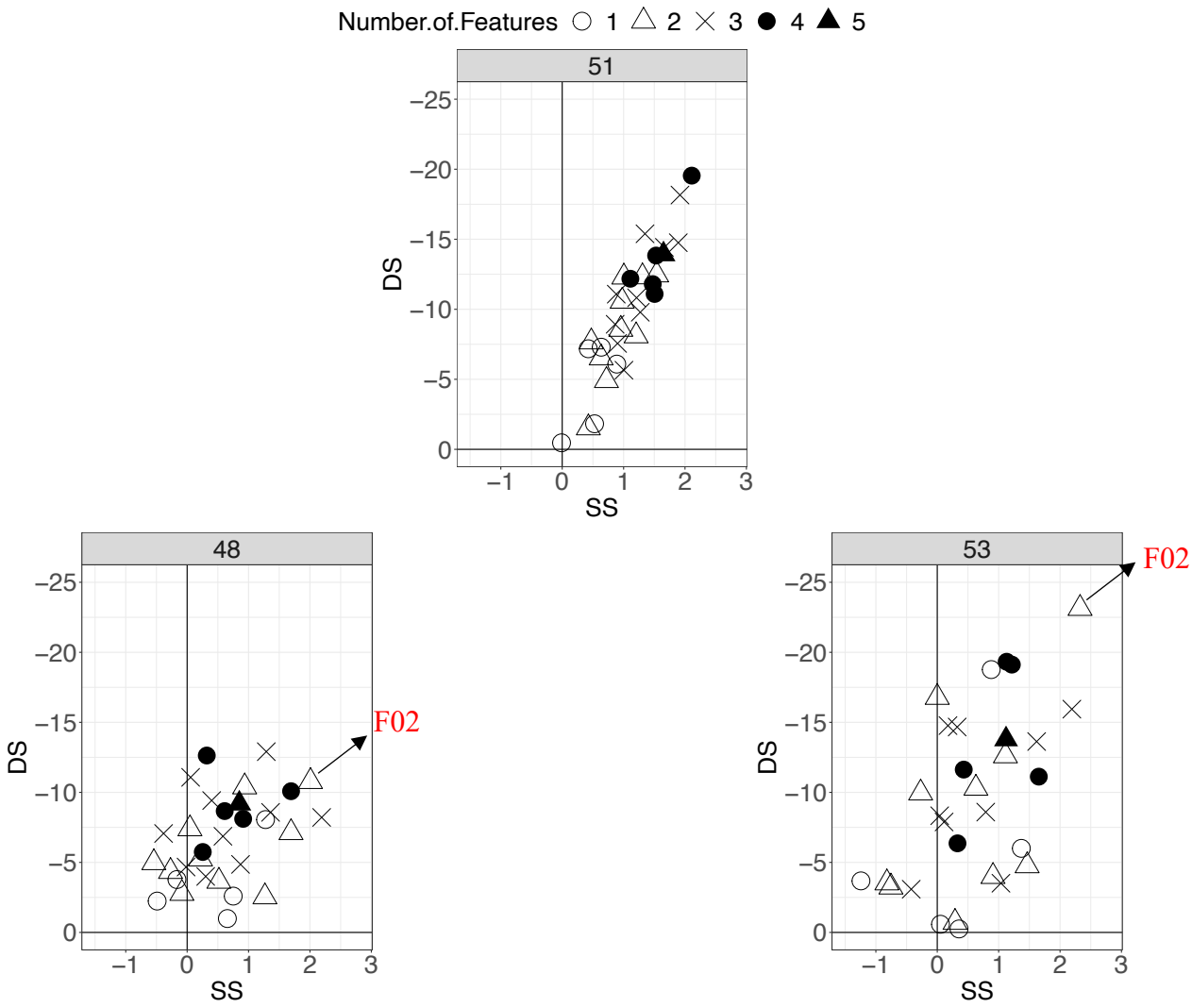


Figure 6. Mean SS and DS LLR of speakers #51, #48 and #53 across 31 systems. Arrows in lower panels indicate that speakers #48 and #53 can be well separated using a two-feature system (F02).

Figure 7 shows the RMSE values in SS (upper panel) and DS (lower panel) comparisons across the 31 systems based on different combinations of input features, indicating the stability in individual speakers' LLR within systems. The black dots represent the RMSE values of each individual speaker, and the coloured triangles are the mean RMSE values of all speakers in each system. The RMSE values were plotted from the highest (left end of the x-axis) to lowest (right end of the x-axis) in terms of the mean RMSE values. Higher RMSE values indicate that

speakers have more fluctuating LLRs relative to their own mean LLR for that system and thus are more sensitive to different configurations of training and reference speakers. Overall, all speakers tend to be more fluctuating in DS comparisons than SS comparisons (note the y-axis limits are different). This is likely due to the fact that speakers can only be so similar to themselves, but infinitely different from each other (Kinoshita et al., 2009). Furthermore, the data used in this study is derived from experimental materials where the stylistic and situational context is very similar, thus limiting the within-speaker variation that might be observed in the real world where individuals speak in very different circumstances and discourse styles. Based on the same LLR scale (Table 1), speakers fluctuate more in SS comparisons when more features are included in the system. However, the fluctuation in individual speakers' SS LLR is relatively small, with all SS RMSE values less than 1 (i.e. within one verbal category; Table 1) across the 31 systems.

For DS comparisons, increasing the number of features does not necessarily lower the stability in individual speakers' mean LLR outputs, e.g. the mean DS RMSE value is higher in the F2 system (ca. 6.25) than in the F0123DUR system (ca. 5). However, among systems with the same number of features, speakers fluctuate more when F2 is involved, i.e. speakers in systems at the left end of x-axis, e.g. the F012, F0123, F0123DUR systems, fluctuate more than those at the right end, e.g. the F1, F1DUR, F3DUR, F3 systems. It is also worth noting that most speakers have DS RMSE values fluctuating between 2.5 and 7.5 in most of the systems (e.g. F2, F12, F123, F0123), indicating that the LLRs of speakers in DS comparisons among 100 sampling replications could be 5 above or below the mean LLR. For example, if one speaker has a mean LLR of -2 in DS comparisons in the F2 system, the possible LLR outputs in the 100 replications could be between -7 and 3.

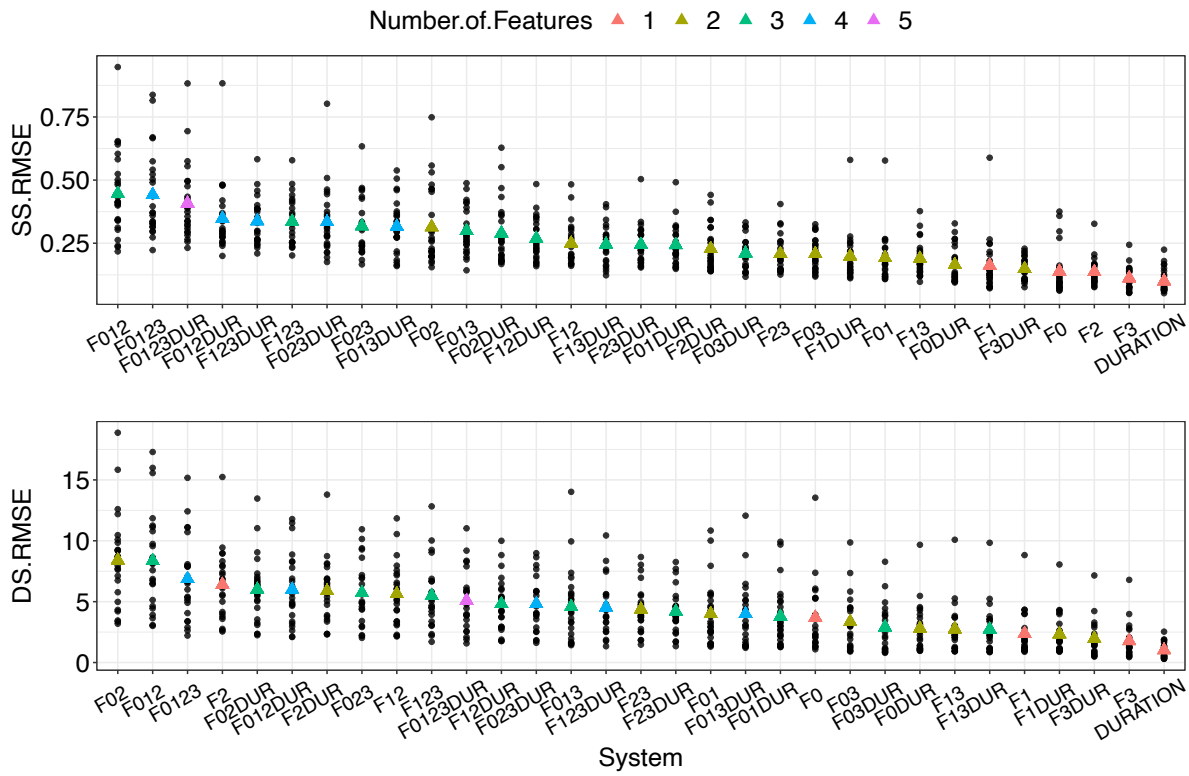


Figure 7. RMSE values of individual speakers across 31 systems.

Figure 8 shows the stability of individual speakers' LLR in systems with the same number of features (the five-feature system is not included as there is only one such system). Each dot represents the difference between the maximum and minimum RMSE values (i.e., the range) of each speaker across systems with equal number of features. In DS comparisons, the majority of speakers tend to be the least stable in two- and three-feature systems and the most stable in systems with one and four features. Around half of the speakers yield DS RMSE values higher than 5 in one-feature systems, while most of the speakers have DS RMSE values lower than 5 in systems with four features. However, speakers show different patterns in SS comparisons. Some speakers start with low SS RMSE values (high stability) in one-feature systems and end up with high SS RMSE values (low stability) in four-feature systems, while some other speakers show an opposite pattern. The remaining speakers show similar patterns to those in DS comparisons where they yield the most fluctuating performance in systems with two or

three features. All the speakers have SS RMSE values lower than 1, indicating that the fluctuation in SS comparisons caused by different combinations of features and configurations of training and reference speakers is lower than one magnitude in terms of strength of evidence.

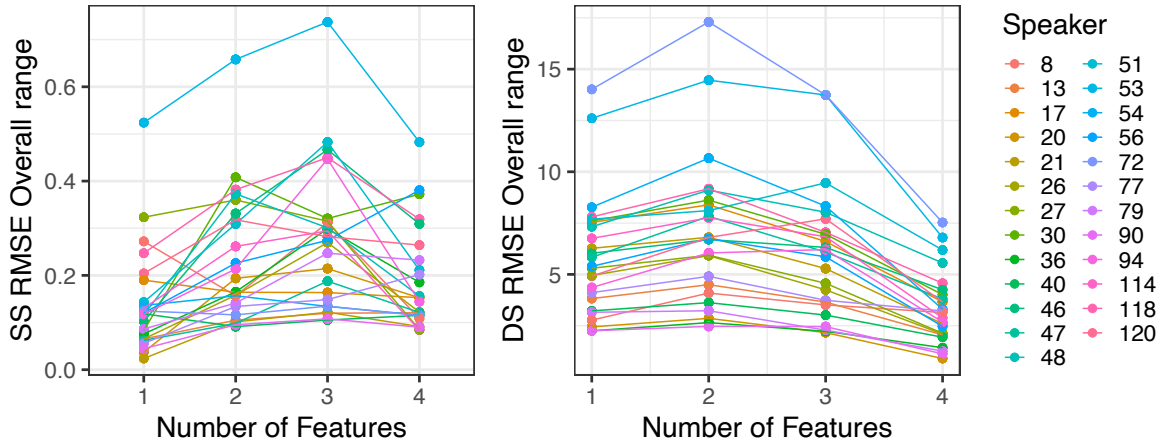


Figure 8. SS and DS RMSE ranges across different systems per test speaker.

*Pearson's r* was calculated to explore the correlation between the mean LLRs and RMSE values for each speaker, i.e. the correlation between the average strength of the evidence and stability of individual speakers' LLR. Table 2 shows the correlation coefficients using individual speakers' SS/DS LLR and RMSE values from systems with the same number of features. Among SS comparisons, the SS LLR shows a positive correlation with RMSE in the five-feature system (*Pearson's r = 0.41*), indicating that speakers who yield higher LLRs (more likely to be separated in SS comparisons) also generally yield more unstable performance. In systems with one to four features, the SS LLR and RMSE values show little correlation. For DS comparisons, the DS LLR and RMSE are negatively correlated in all systems, meaning that speakers who can be well separated from others in DS comparisons also generally produce more unstable performance (note that the DS LLR scale is reversed).

| | SS LLR vs. SS RMSE | DS LLR vs. DS RMSE |
|---|---|---|
| Number of features | *Pearson's r* | *Pearson's r* |
| 1 | 0.16 | -0.73 |
| 2 | 0.14 | -0.52 |
| 3 | 0.14 | -0.62 |
| 4 | 0.13 | -0.58 |
| 5 | 0.41 | -0.69 |

Table 2. Correlation between the validity and stability of individual speakers.

## 4 What happens in a real case?

The results above show that some speakers are more affected by different combinations of features and configurations of training and reference speakers, while others are less affected. Under real case scenarios, there are likely to be limitations on the amount of data that is both available and analysable, especially given the labour-intensive nature of linguistic and phonetic analysis. Therefore, it is realistic to suppose that maximally around 30 training and reference speakers (and in reality, likely fewer) could be sampled from a relevant population (based on sample sizes in many forensic phonetic studies and in a casework report from Rose (2013) who used a total of 35 reference speakers to compute a LR). Although it depends on the specific defence proposition, the real size of the relevant population is likely much larger than the number of training and reference speakers sampled. In order to demonstrate the possible LLR ranges that one can obtain with such samples in a case, the following section shows the LLR ranges from speakers in the system with the best performance in terms of $C_{llr}$ mean and range, i.e. the F0123 system. Based on the SS and DS RMSE values, the speakers who were the most and least fluctuating in SS and DS comparisons were selected. Table 3 shows the RMSE values of the selected speakers from the F0123 system,

| Speaker | SS RMSE | Speaker | DS RMSE |
|---|---|---|---|
| **#40** | 0.22 | **#20** | 2.21 |
| **#30** | 0.84 | **#51** | 15.17 |

Table 3. SS and DS RMSE values of speakers with the least and most variable LLRs.

Figure 9 shows the ranges of SS LLRs of speakers #30 (most fluctuating) and #40 (least fluctuating) across the 100 replications, i.e. #30 and #40 are compared with themselves 100 times with different configurations of training and reference speakers using F0, F1, F2 and F3 as input. The x-axis is the SS LLR values and y-axis shows the speaker IDs. Both speakers yielded positive LLRs, i.e. consistent-with-fact results in all of the 100 replications. Speaker #40 yielded the least variable LLRs, varying from ca. 0.6 to 1.9, which is a difference between *limited* and *moderate support* for SS origin in terms of verbal LLR expression. However, the SS LLR of speaker #30 varies between 1.3 and 6.6, which is a difference between *moderate* and *very strong support* for SS origin. Comparatively, speaker #40 seems to be less problematic, while speaker 30 would be more problematic in a real case, given the much larger uncertainty in the computed LLR values as a function of sampling variability.

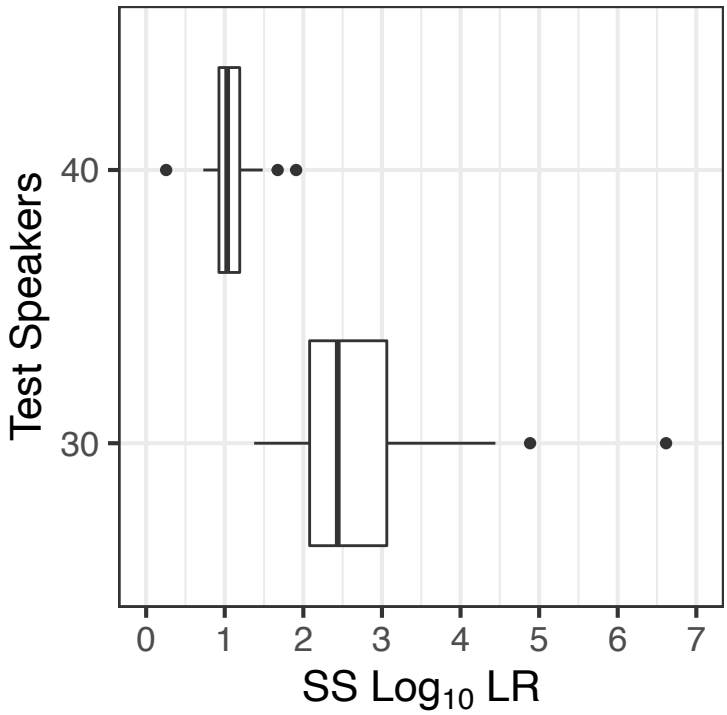


Figure 9. SS LLRs of speaker #30 and #40 using different training and reference speakers in the F0123 system.

Figure 10 shows the DS LLRs of the most and least variable speakers, i.e. speaker #20 and #51, in DS comparisons. The x-axis indicates the DS LLR values, while the y-axis shows the rest of the test speakers that are being compared (NB the x-axis scales are different for top and bottom panels). Each boxplot represents the variation in DS LLRs, e.g. the first boxplot in the top panel indicates the DS LLR ranges of speaker #20 compared with speaker #8, with different configurations of training and reference speakers across 100 replications. It is noted that speaker #20 yields the most stable LLRs in DS comparisons across 100 replications, but this speaker also gives the most contrary-to-fact results when compared with other test speakers. Contrary-to-fact DS LLRs are observed when speaker #20 is compared with 12 (out of 24) speakers (e.g. #17, #21, #26, #27, #30 etc). Among the 12 speakers, speaker #20 is classified as the same speaker as three other speakers in all of the 100 replications, i.e. speakers #27, #54, and #56. Clearly, the contrary-to-fact results for speaker #20 would be misleading under a real

case scenario. The highest DS LLR goes up to 2.12, which indicates a *moderately strong support* for SS origin (given the speech samples are from different speakers). For consistent-with-fact results in speaker #20, the majority of DS LLRs vary between 0 and -5, which is a difference between *limited* and *very strong support* for DS origin. On the other hand, speaker #51 does not to have contrary-to-fact LLRs when compared with the remaining 24 speakers and the DS LLRs range between ca. -12 and -50 for most of the DS comparisons of speaker #51. This pattern indicates that speaker shows weaker strength of evidence also has less variability as well as more contrary-to-fact results.

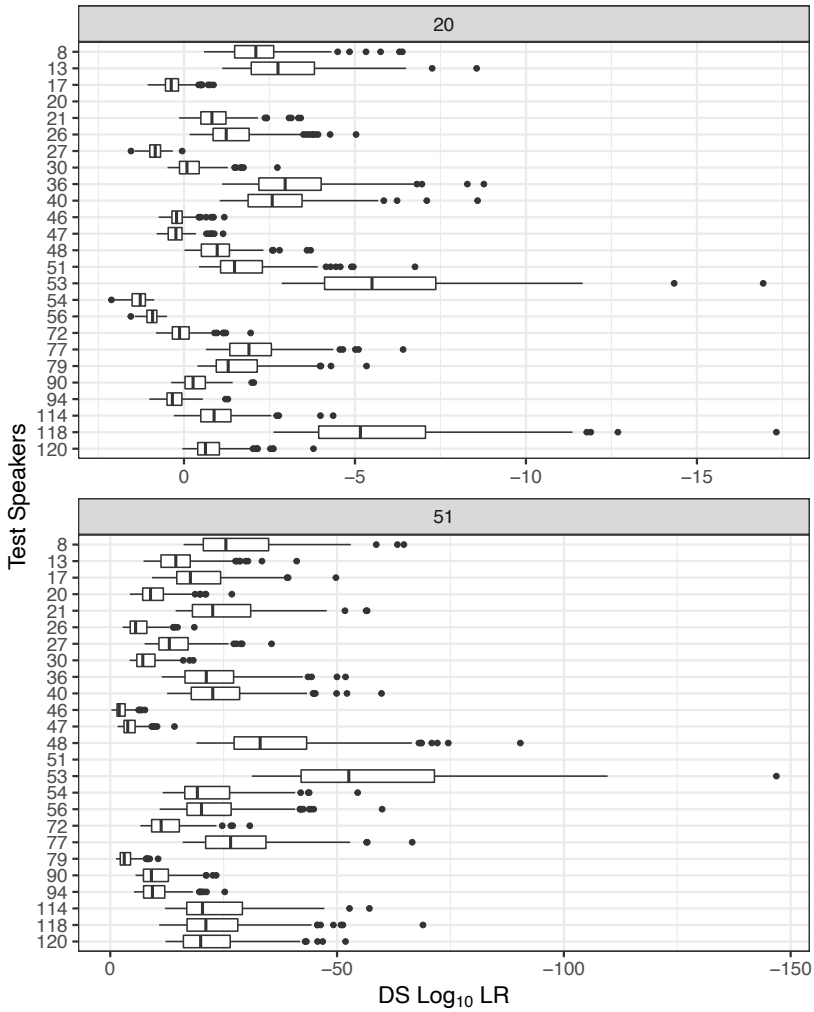


Figure 10. DS LLRs of speaker #20 and #51 using different training and reference speakers in the F0123 system.

# 5 **Discussion**

## 5.1 Summary of general findings

31 systems were evaluated and compared from the perspective of both overall performance and individual speakers' behaviour. In general, $C_{llr}$ mean reduces with more features involved, while no clear pattern is observed in $C_{llr}$ range (Figure 5). Individual speakers display considerable fluctuation.

### 5.1.1 Overall performance

The performance of all 31 systems suggests that systems with more features in general yield better validity. However, in this study there were some exceptions in the F0123 and F0123DUR systems (Figure 4), i.e. the four-feature system outperforms the five-feature system. Therefore, it is important to acknowledge that simply adding extra features does not necessarily improve the overall system validity and stability (and this is especially true when considering individuals). The ranking of different input features also shows a similar pattern to McDougall & Nolan (2007), where F2 yielded the best discriminatory performance, which is inconsistent with other studies showing that higher formants (e.g. F3) have better speaker discriminatory performance. A comparison between the F0123 and F2 systems shows that the F2 system has a lower $C_{llr}$ range (0.04) than that of the F0123 system (0.15). However, all of the $C_{llr}$ values in the F0123 system across 100 replications are lower than those in the F2 system (Figure 4).

### 5.1.2 Individual speakers

In terms of individual speaker behaviour, all speakers yielded mean consistent-with-fact results in systems with four and five features, while contrary-to-fact results were observed in systems with fewer than four features (Figure 6 and Appendix). Furthermore, speakers have different fluctuating patterns in SS and DS comparisons. In SS comparisons, speakers are more likely to yield fluctuating performance in systems with more features (upper panel in Figure 7), indicating that combining multiple features has a chance to increase within-speaker variation, which makes speakers more sensitive to sampling variability. Moreover, Figure 8 shows that speakers have different fluctuating patterns across systems with different numbers of features, e.g. speaker #30 yield the most fluctuating performance in systems with four features, while speaker #51 yield the least fluctuating performance. This suggests that the effect of different combinations of features and configurations of training and reference speakers on the stability of individual speakers' LLR is itself speaker specific. In DS comparisons, there does not seem to be any general relationship between the number of features and individual speaker stability (Figure 7 lower panel). However, when compared in systems with same number of features (Figure 8 right panel), individual speakers become less fluctuating when four features are used, indicating that different combinations of features and configurations of training and reference speakers have less effect on the stability of individual speakers' LLR in DS comparisons. A comparison between the most (speaker #51) and least (#20) fluctuating speakers from the F0123 system in the DS comparisons show that validity and stability of individual speakers' LLR are likely to be negatively correlated. It is worth noting that the results here are necessarily conservative as the random samples of training and reference data used in the current study are not independent of each other across the 100 replications. Therefore, we would expect even

more variability in results if speakers are randomly sampled from other datasets, and truly independent samples are used.

## 5.2 Implications for research

Most previous studies in FVC have tested the speaker-discriminatory performance of linguistic and phonetic features, aiming for lower $C_{llr}$ values. However, the current study shows that it is important to understand and acknowledge the trade-off between system validity (mean $C_{llr}$) and stability ($C_{llr}$ range), e.g., how much variation/uncertainty are we willing to accept given validity? The trade-off between system validity and stability should be case-specific, as each case would have different prerequisites, e.g. the amount of speech available in the suspect and offender samples, the number of speakers available in training and reference data, and how narrowly or broadly defined is the relevant population. Moreover, individual speakers have often been overlooked in FVC research, with focus given to overall performance. The present study shows that while many test speakers will follow the general, group pattern (Figure 6), i.e. more features contribute to better performance, there will necessarily be exceptions (e.g. speaker 53). Thus, considerable variability in LR output may be introduced simply through the typically random selection of speakers to act as training, test and reference data (Figure 9 and 10); variability which is also specific to the choice of input features. Future FVC studies should therefore explore the role of random effects on overall performance and individual speakers' behaviour, rather than reporting $C_{llr}$ with a single configuration of speakers where performance is treated as being representative of the input features and relevant population chosen. It is true that using an automatic speaker recognition (ASR) system might yield better performance when compared with linguistic analysis of a single segment or a group of segments, as is common in linguistic FVC research (although how ASR compares to the linguistic approach

as used in real casework is an empirical question and one that has not been tested adequately). However, a key message we are trying to convey in the current study is the issues of uncertainty and performance variability across individual speakers (section 4).

### 5.3 Implications for casework

In all forms of FVC (auditory-acoustic, semi-automatic or automatic), experts have degrees of freedom to make decisions and conduct analyses, i.e., analysts have to decide which features to be included in an analysis, the relevant population to be used, mathematical models for speaker modelling (e.g., MVKD, Gaussian Mixture Model -Universal Background Model (GMM-UBM) or number of Gaussians to be used for GMM-UBM, Jessen 2021a) and calibration methods (e.g., Morrison & Poh 2018; Wang & Hughes 2021). Those decisions will, in part, be determined by pragmatic considerations, such as whether there are comparable features available for analysis in the two samples. Meanwhile, decisions are also based on the analysts' knowledge of which features carry the greatest speaker discriminatory power (knowledge which is gathered through experience and research, or via a specific validation exercise). For example, the current study has demonstrated how the different choice of linguistic features would affect overall performance and individual speakers' behaviour in relation to sampling variability. Based on the variability observed in the current study, it is our opinion that experts' decisions should be driven by reducing uncertainty in evidence evaluation rather than trying to maximise discrimination or the potential of producing a high validity (i.e., a very low $C_{llr}$ in current context), and that it is crucial to recognise and acknowledge where there is uncertainty within the process. The focus on uncertainty rather than discrimination ultimately reflects the fact that the role of the expert is to aid the trier-of-fact in making better

decisions, which in turn reduces the possibility of miscarriages of justice (Brümmer & Swart, 2014).

Using data-driven approaches allow us to explicitly measure/describe and possibly deal with uncertainty. The data-driven approach, of course, requires data and the implementation of complex mathematical models. Meanwhile, the data-driven approach also involves the issue of explaining and interpreting systems and results to an end-user (e.g., the court). This is not a criticism of the data-driven approach per se, but it is worth acknowledging that this introduces uncertainty in terms of the extent to which the evidence can be reliably used by the trier-of-fact to make better decisions. Conversely, uncertainty does not go away just because the data-driven approach is not employed. In auditory-acoustic analysis (Jessen, 2021b), experts can assess the typicality based on one's expertise or relevant literature. This would result in different degrees of uncertainty in typicality assessments depending on the experts' knowledge/experience or the degree by which data relevant to typicality are available. When it comes to acoustic analysis, different choices of linguistic features and methods, software and settings for measurements would lead to different evaluation results. This has been shown by Roettger (2019) in quantitative-based phonetic studies and in the current study as well. It is therefore important to acknowledge uncertainty and subjectivity, and potentially attempt compensation for that in some way, i.e., attach less weight to a variable when the number of tokens is small. Future work is needed to deal with the issue of uncertainty more systematically.

## 6 **Conclusion**

This study has examined both overall performance and individual speakers' behaviour in LR-based FVC using the first three formants, f0 and vocalic and nasal duration of the FP *um.*

Furthermore, the possible ranges of LLR that one might obtain in real FVC case were also presented and discussed, which provides novel insights for forensic speech science. The ultimate question we have addressed was that whether generalisations drawn from generic system testing would hold for individual speakers under different conditions. Numerous studies (Gwo & Wei, 2016; Morrison, 2016; Morrison & Enzinger, 2016) have discussed the validity and reliability issues in forensic evidence evaluation; however, the variability in LR output observed in the current study has not been addressed in the field of FVC. We propose that overall performance should be tested multiple times using different sets/configurations of training and reference data and the individual speakers' behaviour should be investigated and reported as part of the system testing. Given that all training, test and reference speakers are sampled from the relevant population following certain database selection guidelines (e.g. Morrison et al., 2012), the overall performance and individual speakers' behaviour still vary to different extents. The variability observed in overall performance and individual speakers' behaviour is partially due to statistical issues (e.g. sample size, data extrapolation) and numerous studies have proposed solutions (e.g. different calibration methods) to address this matter (Brümmer & Swart, 2014; Morrison & Poh, 2018; Vergeer et al., 2016). However, a bigger question for forensic phoneticians is that whether there are any systematic linguistic patterns that can be observed to predict or reduce the variability in LR output, e.g. whether a more tailored subset of the relevant population can be selected based on systematic linguistic patterns that would reduce the variability in LR output.

**References**


Aitken, C. G. 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. https://doi.org/10.1046/j.0035-9254.2003.05271.x

Ali, T., Spreeuwers, L., Veldhuis, R., & Meuwly, D. (2015). Sampling variability in forensic likelihood-ratio computation: A simulation study. *Science & Justice*, *55*(6), 499–508. https://doi.org/10.1016/j.scijus.2015.05.003

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2072–2084. https://doi.org/10.1109/TASL.2007.902870

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2–3), 230–275. https://doi.org/10.1016/j.csl.2005.08.001

Brümmer, N., & Swart, A. (2014). Bayesian Calibration for Forensic Evidence Reporting. *Interspeech*, 388–392.

Byrne, C., & Foulkes, P. (2004). The 'Mobile Phone Effect' on vowel formants. *International Journal of Speech Language and the Law*, *11*(1), 83–102.

Champod, C, & Evett, I. W. (2000). Commentary on A.P.A. Breoders (1999) 'Some observations on the use of probability in forensic identification'. *Forensic Linguistics*, *7(2)*, 238–243.

Champod, C, & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, *31*(2–3), 193–203. https://doi.org/10.1016/S0167-6393(99)00078-3

Chen, A., & 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.

Gold, E., & French, P. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech Language and the Law*, *18*(2), 293–307. https://doi.org/10.1558/ijsll.v18i2.293

Gold, E., & French, P. (2019). International practices in forensic speaker comparisons: Second survey. *International Journal of Speech Language and the Law*, *26*(1), 1–20. https://doi.org/10.1558/ijsll.38028

Gwo, C.-Y., & Wei, C.-H. (2016). Shoeprint retrieval: Core point alignment for pattern comparison. *Science & Justice*, *56*(5), 341–350. https://doi.org/10.1016/j.scijus.2016.06.004

Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication*, *94*, 15–29. https://doi.org/10.1016/j.specom.2017.08.005

Hughes, V., Harrison, P., Foulkes, P., French, P., & Gully, A. J. (2019). Effects of formant analysis settings and channel mismatch on semi-automatic forensic voice comparison. *International Congress of Phonetic Sciences.* Melbourne. 3080–3084.

Hughes, V., Wood, S., & 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. https://doi.org/10.1558/ijsll.v23i1.29874

Ishihara, S., & Kinoshita, Y. (2008). How Many Do We Need? Exploration of the Population Size Effect on the Performance of Forensic Speaker Classification. *Interspeech*, Brisbane, 1941–1944.

Jessen, M. (2021a). MAP adaptation characteristics in forensic long-term formant analysis.

*Proc. Interspeech*, 411-415, doi: 10.21437/Interspeech.2021-1697

Jessen, M. (2021b). Speaker profiling and forensic voice comparison. In : M. Coulthard, A. May, and R. Sousa-Silva. (Eds.) *Routledge Handbook for Forensic Linguistics*. 2nd Edition, London: Routledge. pp. 382-399.

Kavanagh, C. (2012). *New Consonantal Acoustic Parameters for Forensic Speaker Comparison* [Unpublished PhD Thesis]. University of York.

Kinoshita, Y., & 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. https://doi.org/10.1558/ijsll.v21i2.191

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *The International Journal of Speech, Language and the Law*, *16*(1), 21. https://doi.org/doi : 10.1558/ijsll.v16i1.91

Kinoshita, Y., Ishihara, S., & Rose, P. (2008). Beyond the Long-term Mean: Exploring the Potential of F0 Distribution Parameters in Traditional Forensic Speaker Recognition. *The Speaker and Language Recognition Workshop*, Odyssey.

Künzel, H. J. (2001). *Beware of the "telephone effect": The influence of telephone transmission on the measurement of formant frequencies', Forensic Linguistics*.

Lennon, R., Plug, L., & Gold, E. (2019). A comparison of multiple speech tempo measures: Intercorrelations and discriminating power. *Proceedings of the 19th International Congress of Phonetic Sciences*, 785–789.

Lo, J. (2018). *FVClrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison* (2.0.1) [Computer software]. https://github.com/justinjhlo/fvclrr

Lo, J. (2021). Seeing the trees in the forest: Diagnosing individual performance in likelihood ratio based forensic voice comparison. *XVII Associazione Italiana Scienza Della Voce Annual Conference*, 34.

Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., Mazzella, W. D., Taroni, F., & Hicks, T. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice*, *56*(5), 364–370. https://doi.org/10.1016/j.scijus.2016.05.009

McDougall, K., & Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. *16th International Congress of Phonetic Sciences*, 1825–1828.

Morrison, G., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W. C., van der Vloed, D., J F Ypma, R., & Zhang, C. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, 61(3), 229-309. https://doi.org/10.1016/j.scijus.2021.02.002

Morrison, G., & Poh, N. (2018). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors. *Science & Justice*, *58*(3), 200–218. https://doi.org/10.1016/j.scijus.2017.12.005

Morrison, G. S. (2009a). Forensic speaker recognition 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. https://doi.org/10.1558/ijsll.v15i2.249

Morrison, G. S. (2009b). Forensic voice comparison and the paradigm shift. *Science & Justice*, *49*(4), 298–308. https://doi.org/10.1016/j.scijus.2009.09.002

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. https://doi.org/10.1016/j.specom.2010.09.005

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. https://doi.org/10.1080/00450618.2012.733025

Morrison, G. S. (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*, *54*(3), 245–256. https://doi.org/10.1016/j.scijus.2013.07.004

Morrison, G. S. (2016). Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Science & Justice*, *56*(5), 371–373. https://doi.org/10.1016/j.scijus.2016.05.002

Morrison, G. S., & Enzinger, E. (2016). What should a forensic practitioner's likelihood ratio be? *Science & Justice*, *56*(5), 374–379. https://doi.org/10.1016/j.scijus.2016.05.007

Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. *The Speaker and Language Recognition Workshop*, Odyssey, 62–77.

Nolan, F. (2001). Speaker identification evidence: Its forms, limitations, and roles. In Proceedings of *Law and language: Prospect and retrospect*. 12-15. Levi, Finnish Lapland.

Nolan, F., McDougall, K., De Jong, G., & 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. https://doi.org/10.1558/ijsll.v16i1.31

R, core team. (2020). *RStudio: Integrated Development for R*. RStudio, Inc. http://www.rstudio.com/

Robertson, B., Vignaux, G. A., & Berger, C. E. H. (2016). *Interpreting evidence: Evaluating forensic science in the courtroom* (Second edition). John Wiley and Sons, Inc.

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*(1), 1. https://doi.org/10.5334/labphon.147

Rose, P. (2003). The Technical Comparison of Forensic Voice Samples. In *Expert Evidence* (Vol. 1–99). Thomson Lawbook Company.

Rose, P. (2013). Where the science ends and the law begins: Likelihood ratio-based forensic voice comparison in a $150 million telephone fraud. *International Journal of Speech Language and the Law*, *20*(2), 277–324. https://doi.org/10.1558/ijsll.v20i2.277

Rose, P. (2011). Forensic voice comparison with secular shibboleths—A hybrid fused gmm-multivariate likelihood ratio-based approach using alveolo-palatal fricative cepstral spectra. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5900–5903. https://doi.org/10.1109/ICASSP.2011.5947704

Rose, P., & Morrison, G. S. (2009). A response to the UK Position Statement on forensic speaker comparison. *International Journal of Speech Language and the Law*, *16*(1), 139–163. https://doi.org/10.1558/ijsll.v16i1.139

Rose, P., & Wang, B. 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*. 326–333. https://doi.org/10.21437/Odyssey.2016-47

Vergeer, P., van Es, A., de Jongh, A., Alberink, I., & Stoel, R. (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? *Science & Justice*, *56*(6), 482–491. https://doi.org/10.1016/j.scijus.2016.06.003

Wang, B. X., Hughes, V., & Foulkes, P. (2019). The effect of speaker sampling in likelihood ratio based forensic voice comparison. *International Journal of Speech Language and the Law*, *26*(1), 97–120. https://doi.org/10.1558/ijsll.38046

Wang, B. X., Hughes, V. (2021) System Performance as a Function of Calibration Methods, Sample Size and Sampling Variability in Likelihood Ratio-Based Forensic Voice Comparison. *Proc. Interspeech*, 381-385, doi: 10.21437/Interspeech.2021-267

Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – Female voices. *Speech Communication*, *55*(6), 796–813. https://doi.org/10.1016/j.specom.2013.01.011

Zhang, C., Morrison, G. S., & Thiruvaran, T. (2011). FORENSIC VOICE COMPARISON USING CHINESE /iau/. *International Congress of Phonetic Sciences, Hong Kong*, 2280–2283.

## Appendix

SS and DS LLR plots of 25 speakers across 31 systems.