DOI: 10.17469/O2108AISV000004

JUSTIN J.H. LO

# Seeing the trees in the forest: Diagnosing individual performance with acoustic data in likelihood ratio based forensic voice comparison

System testing is a crucial part of likelihood ratio based forensic voice comparison, but the evaluation of system performance has thus far focused on global metrics, with little attention paid to the variation in individual performance and the factors behind such variation. Using long-term formant distributions as a case study, this study applies the notion of biometric menagerie to analyse performance on the individual level, and further explores the connection between performance and the underlying acoustic data. Zooplot analysis reveals distinct distributions of how individual speakers perform for each long-term formant. Acoustic analysis further reveals clear patterns in the formant data displayed by speakers with outlying performance. Together, the findings support the view that individual-level analysis can offer useful diagnostic insights into system performance that are unavailable in global-level assessment.

*Keywords*: forensic voice comparison, long-term formant distributions, likelihood ratios, individual performance, zooplots.

## 1. *Introduction*

In cases of forensic voice comparison (FVC), a voice sample from a speaker whose identity is unknown is typically compared with another sample from a known speaker. To this end, features in the voice and speech are analysed to assess both the similarity between the questioned and known samples, and the typicality of the features observed. In recent years, the adoption of likelihood ratios (LRs) in forensic speech science offers an explicit framework to assess voice evidence with respect to two competing hypotheses: (1) that the two speakers are the same, and (2) that the unknown speaker is not the known speaker, but another person in the relevant population. Within this framework, the analyst's conclusions are expressed in the form of a verbal or numerical LR to indicate the direction and degree of support offered by the evidence for one of the propositions. As LRs are sensitive to not only the speech features chosen, but also the method and the database of speakers used in their derivation, these must be tested to ensure that the resultant systems are valid and reliable.

In numerical LR-based FVC, system evaluation is predominantly conducted on the global level, where the strength of performance is indicated by means of a single metric score. The most commonly used metrics are the equal error rate (EER) and

the log likelihood ratio cost function ($C_{llr}$, Brümmer, du Preez, 2006). The EER identifies the percentage of comparisons with contrary-to-fact outcomes, at a score threshold where the rate of misses (same speaker identified as different) is equal to the rate of false matches (different speakers identified as the same). $C_{llr}$, on the other hand, takes into account the magnitude of errors, such that contrary-to-fact LRs of a larger magnitude incur higher penalty and, in turn, higher $C_{llr}$. A $C_{llr}$ that is greater than 1 indicates that the system is poorly calibrated. For both of these metrics, stronger system performance is indicated by smaller values: The closer they are to 0, the better the system is judged to be performing.

Other graphical means of evaluation, such as receiver operating characteristic (ROC) curves, detection error tradeoff (DET) curves and Tippett plots, are also often employed to provide more information about the overall performance of the system (see Morrison, Enzinger, 2016). Commonly used in the assessment of automatic speaker recognition (ASR) systems, ROC curves plot the rate of true matches against the rate of false matches across a range of acceptance thresholds in a single curve, while DET track the rate of false matches against the rate of misses in a similar fashion. Meanwhile, Tippett plots trace performance of same-speaker and different-speaker comparisons in separate curves to visualise the cumulative distribution of scores (or LRs) in each type of comparison.

As much as the above metrics and graphs can illustrate the global level of system performance, their diagnostic value can be limited. A more microscopic view of system performance, beyond rates and sizes of error, can be obtained by examining the performance of individual speakers. By analysing in detail the speakers who perform exceptionally well or disproportionately contribute to errors, researchers can gain insights into the nature of the errors in the system and work towards improving system design in a targeted manner. Further, identifying individual voices who are difficult to match against any speakers in the same database may be helpful for optimising homogeneity within forensic databases (San Segundo, Tsanas & Gómez-Vilda, 2017).

To date, analysis of individual performance remains rarely performed in the context of FVC, and the causes behind speakers being identified as outliers within any tested system are still underexplored. In addition to technical properties of the audio samples, such as non-uniform duration and acoustic quality (Nash, 2019), variation of individual performance in any given system can arise due to physiological and behavioural reasons, as well as the impact of these factors on the quality of data capture (Dunstone, Yager, 2009). Fundamental frequency, for example, is susceptible to variation due to physiological factors in both the short and the long term (Braun, 1985; Rhodes, 2012). In terms of behaviour, speakers may seek to disguise their voice. They may also shout at such a volume that causes clipping in the recording, thus impacting data capture. A close analysis of LRs from individual speakers and the corresponding speech data used to generate those LRs may thus not only be useful in diagnosing the sources of variation in individual

performance, but also more generally help understand the relationship between input and output in numerical LR-based testing.

Within the context of ASR, Alexander, Forth, Nash & Yager (2014) has proposed preliminary links between individual performance and aspects of voice quality. However, subjective judgments of voice quality are far removed from cepstral coefficients, the acoustic features used in ASR systems that are highly abstract, and do not encode the same speaker-specific information (French, Foulkes, Harrison, Hughes, San Segundo & Stevens, 2015). The connection between individual LR performance and the input data thus warrants further investigation.

To address this issue, the present study seeks to extend the analysis of individual LR performance to systems based on linguistic-phonetic variables, using long-term formant distributions (LTFDs) as a case study. It additionally explores the connection between individual performance as derived from LRs and the underlying speech data. Following previous analyses of individual performance, this study makes use of the notion of "biometric menagerie" (see below §3) to classify speakers who perform exceptionally well or poorly. Individual LTF data from these outlying speakers are considered with respect to others in the population to analyse any links with their performance.

## 2. *Long-term formant distributions*

The use of LTFDs as viable speaker discriminants in FVC is first proposed in Nolan, Grigoras (2005). Instead of individual sounds, LTFDs measure the whole collection of formant estimates from all vowels (or voiced sounds) present in the recording. LTFDs are thus argued to capture the overall filtering behaviour of the supralaryngeal vocal tract, reflecting not only its physiology but also the speaker's idiosyncratic articulatory habits. There is empirical evidence in support of the latter, as LTF1 and LTF2 means have respectively been found to correlate with the vocal settings of raised/lowered larynx and fronted tongue body (French et al., 2015). Further support for the inclusion of LTFDs in the FVC toolkit can be found in their independence from other acoustic and temporal measures, such as fundamental frequency and speaking rate (Moos, 2010).

Speaker-specific information in LTFDs lies not only in the location of the peaks or means of the distributions, but also in their overall shapes (Cho, Munro, 2017). Studies conducted within the LR framework have tested the discriminatory potential of LTFDs in English and German and reported low error rates in both languages (Becker, Jessen & Grigoras, 2008; French et al., 2015; Gold, French & Harrison, 2013). LTFs of higher formants have been found to provide stronger performance than those of lower formants (Gold et al., 2013), while using multiple formants in combination can further improve performance (Becker et al., 2008; Gold et al., 2013).

As semi-automatic formant-based acoustic features, LTFDs are readily interpretable in linguistic-phonetic terms, as opposed to cepstrum-based features used in automatic approaches, which do not display any comparable degree of interpretability (Rose,

2003). This characteristic is considered to make LTFDs a useful set of features for the present exercise. Previous attempts at individual analysis for LTFDs have also sought to explore their links with individual LR performance, such as in Hughes, Harrison, Foulkes, French, Kavanagh & San Segundo (2018), who found considerably variable behaviour across individual speakers, not only in the strength of evidence produced, but also in the stability of their performance in face of channel variation. Importantly, in systems that combined LTF1-3, their corresponding bandwidths and delta coefficients as input, none of LTF1-3 means was found to be able to predict an individual's performance, although it must be noted that only the relationship between performance and means was considered but not that between performance and the speakers' full distributions.

## 3. *The biometric menagerie*

The current study makes use of zooplots, a diagnostic tool used in the field of biometrics to visualise individual user performance. The zooplot is built upon the idea of a "biometric menagerie", developed by Doddington, Liggett, Martin, Przybocki & Reynolds (1998) and later expanded on by Dunstone, Yager (2009), where speakers are classified into user groups or animals based on their individual performance. In a zooplot, each speaker's performance in same-speaker comparisons (or genuine performance) is plotted against their performance in different-speaker comparisons (or imposter performance). The use of zooplots thus facilitates the identification of problematic speakers in the database for further analysis and diagnosis. Additionally, the distribution of speakers in a zooplot can be indicative of systematic weaknesses in the algorithm or the database used. A predominance of speakers who perform well in different-speaker comparisons but poorly in same-speaker comparisons, for example, may be an outcome of poor-quality enrolment in the database (Dunstone, Yager, 2009).

In their original formulation, Doddington et al. (1998) distinguishes a default group of speakers, *sheep*, from other speakers who tend to contribute disproportionately to system errors. These animal groups include *goats*, whose voices are particularly difficult to match and hence likely produce errors in same-speaker comparisons; *lambs*, who may disproportionately account for false matches due to their voices being easily imitable; and *wolves*, whose voices may easily imitate others' and thus also contribute to false matches. Dunstone, Yager (2009) introduces a set of relational animals, which are defined by the relationship *between* a speaker's performance in same-speaker comparisons and in different-speaker comparisons, rather than their performance in a single type of comparisons. Described in FVC terms, these groups include:

- *doves*, who perform relatively well in both types of comparisons;
- *worms*, who perform relatively poorly in both types of comparisons;
- *phantoms*, whose voice characteristics are difficult to match against any speaker and so who perform well in different-speaker comparisons but poorly in same-speaker comparisons; and

– *chameleons*, who can be easily matched with (or camouflage as) any speaker and thus perform well in same-speaker comparisons but not in different-speaker comparisons.

This study focuses on the relational groups introduced by Dunstone, Yager (2009), as they allow specific speakers who belong to each of these groups to be identified. Members of each group can then be further analysed for the causes behind their outlying performance within the tested system to be potentially diagnosed.

## 4. *Methodology*

### 4.1 Materials

In the present study, a subset of high-quality microphone recordings from the *Voice ID Database* (Royal Canadian Mounted Police, 2010-2016) were used, consisting of 60 adult male bilingual speakers of Canadian English and French (mean age = 27.7 years), all of whom recorded the same sets of read materials in both languages. Metadata on speakers' language and social background were limited, and speakers were selected on the basis that they participated in recording for both languages and did not self-report knowledge of any other languages. 23 and 31 speakers reported their first language to be English and French respectively (mean age of first exposure to second language = 5.3 years), while the remainder reported to be simultaneously bilingual in both. The current analysis is limited to the data from English.

All speakers were recorded reading 20 phonetically balanced sentences extracted from the Harvard sentences. 22 speakers were also recorded reading an abridged version of *The Rainbow Passage* (Fairbanks, 1960). All recordings were first orthographically transcribed in *Praat* (Boersma, Weenink, 2016), where hesitations, repetitions, mispronunciations and deviations from the set material were retained as far as possible, although partial words were ignored. Automatic segmentation was performed using the *Montreal Forced Aligner* (McAuliffe, Socolof, Mihuc, Wagner & Sonderegger, 2017), which was then manually checked for errors in alignment and corrected where necessary.

While the quality of the recordings may result in validity measures that are more optimistic than those obtained from forensically realistic materials, the controlled nature of the materials has the advantage of ensuring that variation in LTFDs can be attributed to speaker physiology and behaviour, rather than differences in speech content.

### 4.2 Data extraction

Formant centre frequency estimates for the first four formants were extracted in *Praat* at 10 ms intervals from the onset to the offset of all vowels and glides. Formant extraction was automated with a *Praat* script, using the Burg algorithm for linear predictive coding, with the formant tracker set to search for 6 formants up to a maximum formant frequency of 5500 Hz in 25 ms frames. These settings were determined by preliminary testing to be the most appropriate for the current set

of speakers and remained fixed for all speakers. Each speaker's recording contains 26.2 s of pure vowels and glides on average, yielding a grand total of 172,054 sets of formant estimates from all speakers for system testing.

### 4.3 Performance testing

Five systems, each using a different set of input parameters, were tested in total. In addition to the combination of all four LTFs, each LTF was tested individually to facilitate one-to-one comparison with the acoustic data.

To test each system, the 60 speakers were randomly divided into *test*, *training* and *background* sets, each made up of 20 speakers. The computation of LRs followed the two-stage process set out in Morrison (2013). In the *feature-to-score* stage, comparisons were conducted for all speaker-pairs in the *test* set, with each speaker acting in turn as the questioned and known speaker for all speakers, resulting in 20 same-speaker and 380 different-speaker comparisons. LTFDs were modelled and compared using the Gaussian mixed model-Universal background model (GMM-UBM) approach (Becker et al., 2008; Reynolds, Quatieri & Dunn, 2000), implemented by means of the *mclust* package in R (R Core Team, 2018; Scrucca, Fop, Murphy & Raftery, 2016). As only a single recording was available from each speaker, the first half of the recording was taken to form the known sample for the speaker, and the second half was taken to form the questioned sample. The reference population was modelled with a UBM, a single GMM composed of 12 Gaussians calculated using data pooled together from all 20 speakers in the *background* set. In each comparison, a GMM for the known speaker was derived by using data from the known sample to adapt the UBM by means of maximum a posteriori estimation, and the output score of the comparison was calculated by Equation (1). In the *score-to-LR* stage, the *test* scores were then calibrated with output scores similarly computed from the *training* set to obtain $\log_{10}$LRs (LLRs) in a logistic regression procedure. LLR > 0 in a comparison indicates support for the same-speaker hypothesis, while LLR < 0 indicates support for the different-speaker hypothesis. As such, negative LLRs in same-speaker comparisons and positive LLRs in different-speaker comparisons are considered to be contrary-to-fact errors.

$$\text{Score} = \frac{1}{N}\sum_{i=1}^{N} \log_{10} \frac{p(x_i|S)}{p(x_i|B)} \quad (1)$$

where $x_1, x_2, ..., x_N$ correspond to each set of input data from the questioned sample, $S$ = suspect GMM and $B$ = background UBM.

System validity was then assessed using EER and $C_{llr}$. The whole sampling and testing procedure was repeated 100 times, in order to minimise the effect of speaker random sampling on the resultant LRs and validity metrics (Wang, Hughes & Foulkes, 2019), as well as to ensure all speaker-pairs were compared.

### 4.4 Individual-level analysis

To visualise individual LR performance, zooplots for each system were constructed by plotting each speaker's average performance in different-speaker comparisons against their performance in same-speaker comparisons, where speakers with stronger performance were positioned towards the top and the right of the plot. In this study, average performance of any speaker is defined as the arithmetic mean of LLRs from all same- or different-speaker comparisons across 100 repetitions involving that speaker.

Individual-level analysis of LR performance was conducted in three ways. First, the overall distribution of speakers on the zooplots was analysed, in terms of both the absolute values of LLRs (as they can be directly interpreted) and the relative positions of speakers. Second, speakers with outlying performance in each system were identified and categorised into one of the four relational animal groups defined in §3. Following Dunstone, Yager (2009), *doves* are defined as speakers whose average performance is within the best 25% of all speakers in both types of comparisons. This group can therefore be located in the top right corner of the zooplot, as they comprise speakers with the *highest*, most positive mean LLR in same-speaker comparisons (SS-LLR) and the *lowest*, most negative mean LLR in different-speaker comparisons (DS-LLR). Worms, phantoms and chameleons are analogously defined, as outlined in Tab. 1. Speakers whose average mean LLR lies between the top (or bottom) 25% to 30% were additionally identified as *near*-members of animal groups to mitigate cliff-edge effects of borderline cases documented in O'Connor, Elliott, Sutton & Dyrenfurth (2015). Membership of relational groups was then analysed in each system and compared across different systems. Third, to explore the relationship between LR performance and acoustic data, LTFDs of speakers classified as (*near*-)members of relational groups were compared with those of the other speakers.

Table 1 - *Inclusion criteria for doves, worms, phantoms and chameleons and their locations in zooplots*

| | **Mean same-speaker LLR** | **Mean different-speaker LLR** | **Location** |
|---|---|---|---|
| *Doves* | Highest 25% | Lowest 25% | Top right |
| *Worms* | Lowest 25% | Highest 25% | Bottom left |
| *Phantoms* | Lowest 25% | Lowest 25% | Top left |
| *Chameleons* | Highest 25% | Highest 25% | Bottom right |

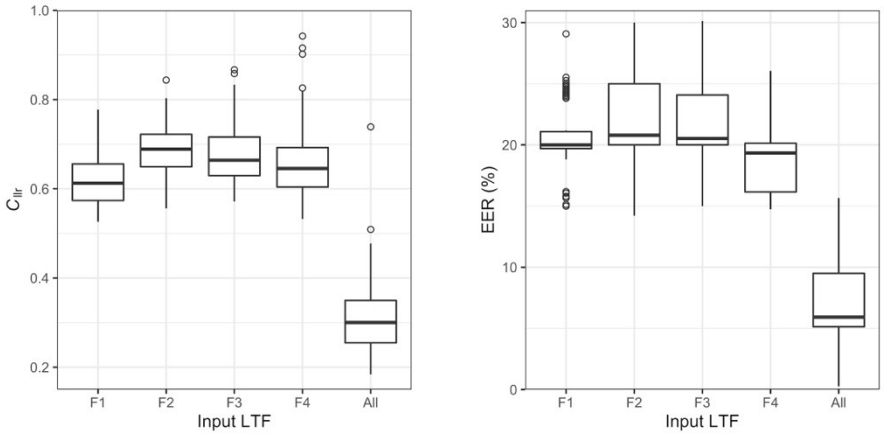
## 5. *Results*

### 5.1 Global metrics

Performance for each system, measured in $C_{llr}$ and EER, is illustrated in Fig. 1. Mean $C_{llr}$ for all systems (and indeed all repetitions) was below 1, indicating that

the systems were well calibrated. The system combining all four LTFs reported the lowest mean $C_{llr}$ and EER at 0.31 and 7.3%, clearly outperforming systems using only single LTFs. Individually, each of LTF1-4 performed at similar levels, with a mean $C_{llr}$ between 0.6 and 0.7 and a mean EER of around 20%. Nevertheless, out of all systems using individual LTFs, Fig.1 demonstrates a trend of LTF2 performing the worst, with the highest mean $C_{llr}$ (0.69) and EER (22.1%), while LTF1 had the lowest mean $C_{llr}$ (0.62) and LTF4 had the lowest mean EER (19.0%).

Figure 1 - *Boxplots of $C_{llr}$ (left) and EER (right) from system testing*




5.2 Individual-level analysis

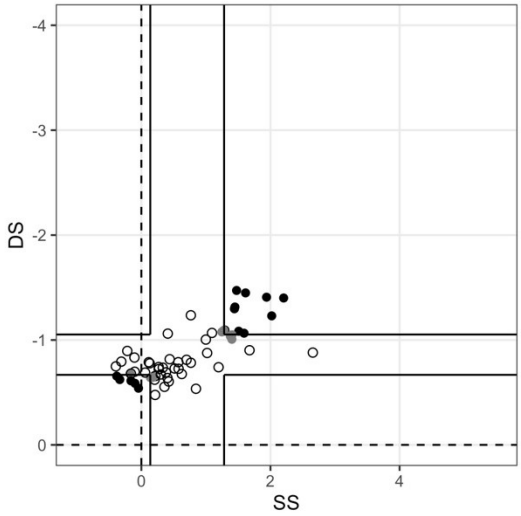
This section presents results on individual LR performance in the tested systems. In each system, the distribution of LLRs is described with the visual aid of a zooplot. The scales of all zooplots are fixed, in order to facilitate direct comparison across different systems. This is accompanied by an analysis of the classification of animal groups in each system and followed by a comparison of group membership across all systems.

5.2.1 LTF1

The zooplot shown in Fig. 2 illustrates the individual LR performance of all 60 speakers in the system based on LTF1. As demonstrated in the zooplot, all speakers reported a negative mean DS-LLR. While mean SS-LLR was positive for most speakers, 12 speakers (20%) reported a negative mean SS-LLR, suggesting that they were not well matched with themselves on average. Performance in SS and DS comparisons is strongly correlated ($|r| = .74$, $p < .0001$): Speakers with stronger performance in SS comparisons similarly reported stronger performance in DS comparisons. This is further supported by the absence of any *phantoms* or *chameleons*, as only members of the other two relational groups (10 *doves* and five *worms*) were identified. Overall, the system produced a narrow range of mean LLRs, with SS-LLR between -0.40 and 2.55, and DS-LLR between -0.48 and

-1.47. Most speakers could be found in a dense cluster near the lower left corner of the zooplot.

Figure 2 - *Zooplot for system with LTF1 as input (abscissa and ordinate respectively show mean SS-LLR and DS-LLR in $log_{10}$LR; solid line segments represent 25th and 75th percentiles; dotted lines indicate mean LLR = 0; members and near-members of relational groups respectively in black and grey)*




5.2.2 LTF2

As shown in the zooplot in Fig. 3, all speakers also reported a negative mean DS-LLR in the LTF2 system, and a majority of speakers reported a positive mean SS-LLR. Although there were the same number of speakers with negative mean SS-LLR (12) as in the case of LTF1, the magnitude of these negative SS-LLRs was slightly higher (up to -0.80), indicating poorer performance in SS comparisons. Performance in SS and DS comparisons is only moderately correlated ($|r| = .51$, $p < .0001$), with a notable presence of speakers in the upper left region of the zooplot, including three speakers who were identified as *phantoms*. By contrast, only nine *doves* and three *worms* were identified for LTF2, fewer than any other individual LTFs.

Figure 3 - *Zooplot for system with LTF2 as input*

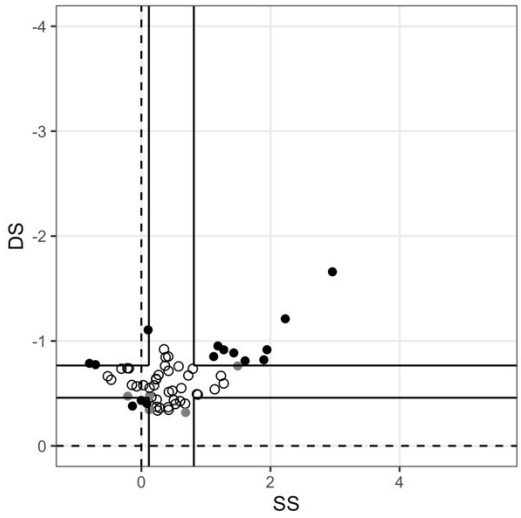
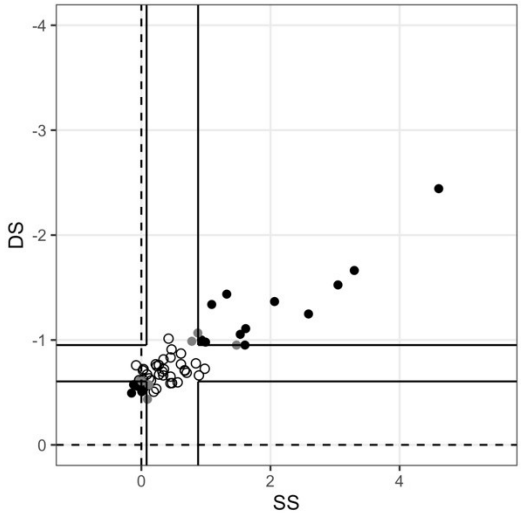



Figure 4 - *Zooplot for system with LTF3 as input*




5.2.3 LTF3

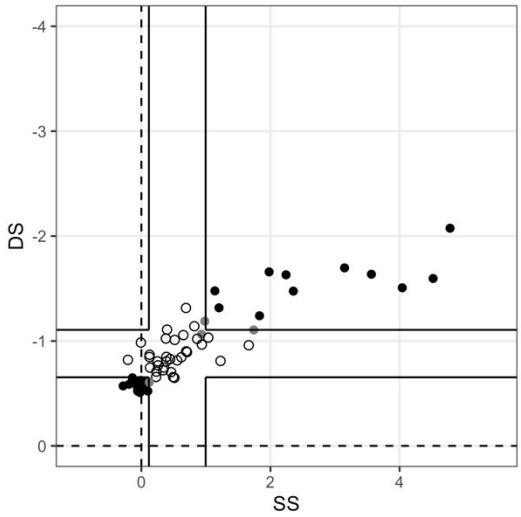
The distribution of speakers in the zooplot for LTF3, as displayed in Fig. 4, shows broad similarities with that for LTF1. No speakers were classified as *phantoms* or *chameleons*, and a strong positive correlation was similarly found between mean SS-LLR and DS-LLR ($|r| = .92, p < .0001$). As in the case of LTF1, a dense cluster of speakers can be located near the lower left corner of the zooplot, including eight *worm* speakers. Despite the relatively high number of *worms*, only eight speakers

reported a negative mean SS-LLR, none of which exceeded -0.15, indicating smaller contrary-to-fact LLRs on average. Stronger individual performance is also evident among the best-performing *doves* (12 members and 3 *near*-members), who are distant from the dense cluster of speakers and are further spread upward and rightward in the zooplot, as a result of more positive mean SS-LLRs and more negative mean DS-LLRs.

### 5.2.4 LTF4

The zooplot for the LTF4 system, as shown in Fig. 5, illustrates a distribution of speakers resembling that for LTF3, although speakers are less clustered towards the lower left corner, indicating an overall greater degree of between-speaker variation in LR performance. Performance between SS and DS comparisons is also strongly correlated ($|r| = .87$, $p < .0001$), with particularly high mean SS-LLR (up to 4.78) reported for a number of speakers. At the same time, this system reported the highest number of speakers with outlying performance, where a total of 11 speakers were classified as *doves* and 13 were classified as *worms*. The high proportion of outlying speakers suggests that, in the case of LTF4, individual performance is more extremely distributed.

Figure 5 - *Zooplot for system with LTF4 as input*

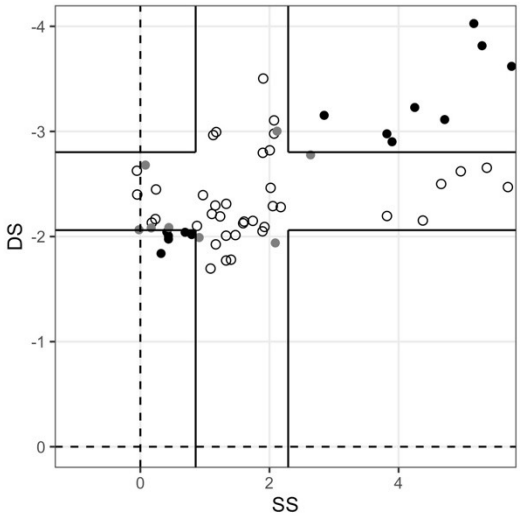



### 5.2.5 All LTFs combined

Compared to the zooplots for individual LTFs, there is evidently a wholesale shift of speakers towards the top of the zooplot in Fig. 6, brought on by more highly negative DS-LLR. Speakers also tend to have much higher SS-LLR, in contrast with zooplots for the other systems, as evidenced by the rightward spread of speakers in Fig. 6. The majority of speakers (42) reported a mean SS-LLR greater than 1, while only three speakers reported marginally negative mean SS-LLR (up

to -0.05). In this system, mean SS-LLR and DS-LLR is moderately correlated ($|r| = .59, p < .0001$). No *phantoms* or *chameleons* were found (although one speaker was classified as *near-phantom*), while eight *doves* and six *worms* were identified. Overall, the distribution of speakers in Fig. 6 is clearly much less clustered than in any of the individual LTF systems.

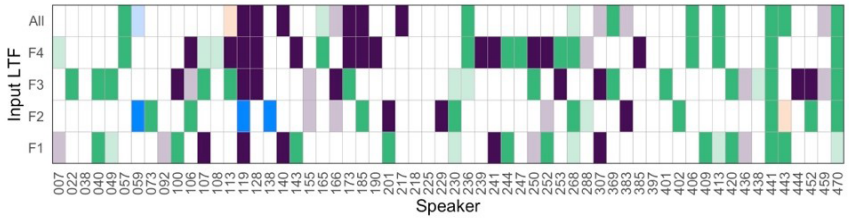
Figure 6 - *Zooplot for system with all four LTFs as input*




5.2.6 Speaker classification

Following separate analysis within each system above, this section considers the performance of individuals across all LTFD systems. Fig. 7 summarises the animal group classification of all speakers for each system tested. While there are some overlaps between different LTFs, it is clear that each LTF generally captured a different group of outlying speakers. Out of 60 speakers, 21 (35%) were in or near the same group for more than one LTF, but only three speakers (5%) were in or near a relational group for all four LTFs: 441 was consistently classified as a *dove*; 470 was similarly always in or near the *dove* group; 119 was classified as a *worm* for all LTFs except for LTF2, where he was classified as a phantom. The difference between systems is further illustrated by the finding that 12 (20%) speakers were classified as the best-performing *doves* for one LTF but as the worst-performing *worms* for another. Across all four LTFs, only five speakers (8%) were not in or near any group, meaning that the vast majority of speakers could be considered an outlying speaker for at least one LTF.

Figure 7 - *Summary of speakers classified as doves (green), worms (purple), phantoms (blue) or chameleons (orange) in all systems tested (near-members of each group in lighter shade)*



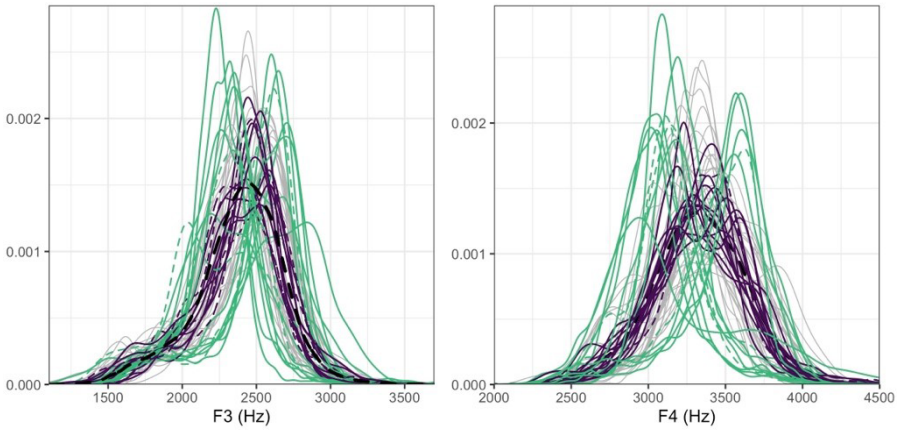

In the combined system (top row, Fig. 7), all but two speakers who were in or near an animal group were similarly classified for at least one of the individual LTFs. Out of those 22 speakers, 12 of them shared their classification with the LTF4 system. While systems using other LTFs shared fewer (*near-*)members of animal groups with the combined system, with LTF2 having only five speakers in common, each of the individual LTFs could uniquely account for at least one classified animal in the combined system. The only exceptions were 217, who was a *worm* in the combined system but not a member of the outlying groups for any individual LTF, and 113, who was a *near-chameleon* in the combined system. The latter could be explained by the mixed contribution of his relatively poor DS performance in LTF4, for which he was classified as a *worm*, and relatively excellent SS performance in LTF3, which resulted in a *dove* classification.

Overall, the analysis here provides evidence on an individual level that complementary speaker-specific information is available from LTFDs of different formants. As the findings above illustrate, speaker-specific information from each LTFD contributes, albeit in varying degrees, to the classification in the combined system.

### 5.3 Acoustic analysis

Results of acoustic analysis from the higher formants are first presented here. The full distributions of each speaker's LTF3 and LTF4, as displayed in Fig. 8, show considerable between-speaker variation in the location of the peak frequency. Peaks in individual LTF3 distributions range from just above 2000 Hz to nearly 2900 Hz, while peaks in LTF4 distributions range from below 3000 Hz to around 3600 Hz. A high degree of variability is similarly found in the shape of the distribution, in terms of both the height and the sharpness of the peaks. Bimodal distributions are also evidenced, albeit rarely, for individual speakers.

Figure 8 - *Distributions of LTF3 (left) and LTF4 for all speakers, with doves in green, worms in purple and other speakers in grey (dashed coloured lines indicate near-members of animal groups; black dashed line indicates group distribution pooled from all 60 speakers)*




A close examination of the speakers classified as *doves* in LTF3 and LTF4 suggests that these are all speakers with relatively extreme peak frequencies, on both the low and high ends of the collection of LTFD. The *doves* in Fig. 8 show a clear separation from the *worms* (and *near-worms*), who are generally close to the group norm, represented by the pooled distribution. The distributions of the *worm* speakers show peak frequencies that are much further away from the extremes. Their shapes are also relatively unremarkable, with no particularly sharp peaks.

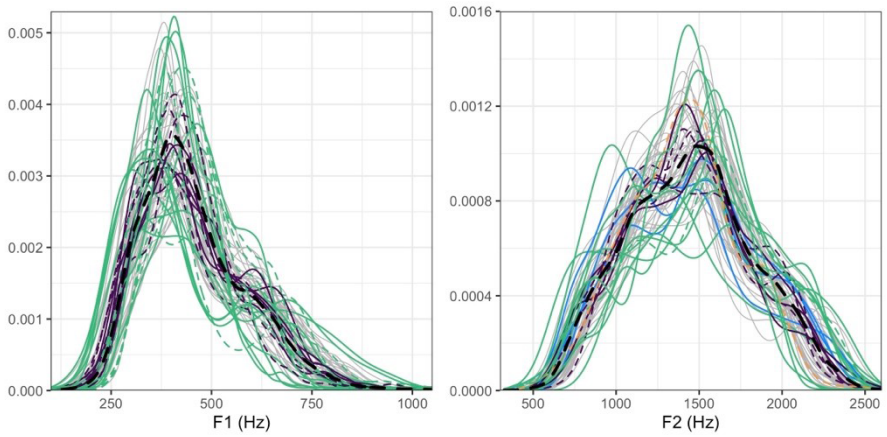
Moving to the lower formants, Fig. 9 shows that LTF1 peaks are limited to a narrow range with relatively little between-speaker variation, which is not surprising in the present context, as F1 is correlated with vowel height and so would be highly constrained when the speech material is uniform across speakers. Nevertheless, the shape of the LTF1 distributions demonstrates substantial variability, especially within the region of frequencies above 500 Hz, where secondary peaks can be found for numerous speakers.

Fig. 9 further shows considerable variability in the distributions of LTF2, especially in the shapes of the distributions. While the peaks for most speakers reside between 1400 and 1600 Hz, the presence of secondary peaks and bimodality is not uncommon among this group of speakers.

As in the cases of LTF3 and LTF4, speakers classified as *worms* (and to a lesser extent, *near-worms*) show distributions that strongly resemble the group norms, both in LTF1 and LTF2. Similarly, many *dove* speakers can be identified as those whose distributions show peaks at especially low or high frequencies within this population. However, it is also clear that the distributions of some *doves* have peaks at frequencies that are by no means extreme, but indistinguishable from the group average. The distributions of these speakers nonetheless show distinctiveness in other ways, through either particularly sharp peaks (in LTF1 and LTF2) or bimodality (in LTF2). There

are also three *phantoms* and one *near-chameleon* in the LTF2 system. Due to their low counts, the acoustic correlates of these groups are not analysed in detail, although it can be noted that the three *phantoms* all appear to demonstrate strong bimodality.

Figure 9 - *Distributions of LTF1 (left) and LTF2 for all speakers, with doves in green, worms in purple, phantoms in blue, chameleons in orange and other speakers in grey (dashed coloured lines indicate near-members of animal groups; black dashed line indicates group distribution pooled from all 60 speakers)*




In summary, the acoustic analysis presented in this section demonstrates a clear contrast between *doves* and *worms* in their LTFDs. *Dove* speakers are mostly accounted for by peaks of relatively extreme frequencies. This is most clearly demonstrated in LTF3 and LTF4, but can also be found in the lower formants. While the remaining *doves* display unremarkable peak frequencies, they show other distinctive characteristics in their distributions. By contrast, the distributions of speakers classified as *worms* all tend to very similar to the overall distributions of the group, in terms of both peak frequency and shape.

## 6. *Discussion*

The current study set out with two main aims, namely to investigate the effectiveness of individual-level analysis in LR-based FVC, and to explore the relationship between individual LR performance and the underlying data from semi-automatic linguistic-phonetic variables. This section discusses these two themes in turn.

Results from system testing of LTFDs show that, in terms of both $C_{llr}$ and EER, LTFDs all performed at similar levels when tested individually but reported much stronger system performance when tested in combination. These results corroborate previous findings of higher discriminatory power in formant-based systems using a combination of multiple parameters (Becker et al., 2008; Gold et al., 2013; Hughes, Wood & Foulkes, 2016), as speaker-specific information from each parameter is

modelled in conjunction. Zooplot analysis undertaken in the current study further illustrates the complementarity of speaker specificity from different LTFDs on the individual level, demonstrating how they capture very different groups of speakers with outlying performance.

In the present study, while the worst performance out of all LTFDs was obtained from LTF2, LTF1 reported $C_{llr}$ and EER that were at least on par with, if not marginally better than, LTF3 and LTF4. Earlier findings that higher formants outperform lower formants in LTFDs (Gold, et al., 2013) are thus not borne out here. Indeed, studies examining the discriminatory potentials of vowel formants have not yielded consistent findings as to whether higher formants convey a greater amount of speaker-discriminatory information than lower formants, which McDougall (2004) argues may depend on the speech materials and conditions. At the same time, when speaker classification is compared across systems, the influence of higher formants appears to dominate in the combined system, as evidenced by the high proportion of classifications shared with LTF3 and LTF4, whereas the contribution of the lower formants is comparatively limited. Zooplots here showed that, in the cases of LTF3 and LTF4, although most speakers had SS-LLRs of low magnitude that were clustered near 0, there were a number of (*dove*) speakers with exceptional mean LLRs, such that they were distant from the other speakers. LTF1, on the other hand, had no such speakers, and indeed reported a narrow range of mean SS-LLRs and DS-LLRs overall. Thus, it may be the case that the performance of higher formants here was driven by a small number of individuals who performed exceptionally well and thus carried over to the combined system, whereas the performance of LTF1 was driven by the population as a whole.

Beyond classification, the zooplots also allow identification of speakers who contribute most to errors in a system. Among the systems tested, mean SS-LLRs and DS-LLRs are well correlated, but stronger performance is generally found for different-speaker comparisons than for same-speaker comparisons, with a number of speakers reporting negative mean SS-LLRs on average. While other graphical means commonly used to illustrate system performance, such as Tippett plots, can also visualise the difference between the two types of comparison and the proportion of contrary-to-fact LRs, in a zooplot analysis the individuals can be placed into focus for further analysis.

Turning to the relationship between individual performance and the underlying acoustic data, in the case of LTFDs, a close correspondence between distinctive distributions and *dove* membership was found across all four formants. For many *dove* speakers, their LTFD peaks lie on the margins of the population distributions, which in turn lead to more skewed LTFDs. This was especially the case in LTF3 and LTF4, where there is considerable between-speaker variability in the location of peaks within this population. Distributions of the remaining *dove* speakers are distinctive in other ways, such as particularly sharp peaks or bimodality. *Worm* speakers, on the other hand, had LTFDs characterised by their proximity to the population norms. These findings may in part explain why LTF1-3 means could

not predict animal group classification in Hughes et al. (2018), who explored this relationship using linear regression. Within individual LTFDs, speakers with very low or very high means can both be considered atypical, whereas it is speakers with means near the middle – not near the lower end – of the group who are considered highly typical and tend to provide weak evidence. The current study differs from Hughes et al. (2018) in that, in the present analysis, the underlying LTFDs were only compared with systems using individual formants, and not the combined system as was the case in Hughes et al. (2018), who also included formant bandwidths and delta coefficients in their systems. Therefore, other factors likely also contribute to the lack of association between the acoustic data and speaker performance. Nevertheless, it remains that the relationship between LTFD means and LR performance cannot be captured linearly.

The present study may have benefited from the uniform and high quality of the recordings, where confounding technical factors are not present and more effective harnessing of speaker-specific information in LTFDs is made possible. In forensically realistic materials, the utility of vowel formants may be limited due to effects of telephone transmission, especially on F1 (Byrne, Foulkes, 2014), and channel mismatch, leading to poorer and more unstable LR performance (Hughes et al., 2018), in which case individual analysis may be even more illuminating as to the factors that drive the performance of individual speakers. Further limitations include the use of controlled materials, which have enabled the isolation of the factor of speakers when examining individual variation in LTFDs but may yield LTFDs different from those derived from spontaneous speech due to the effect of style (Moos, 2010). Future investigations making use of poorer-quality data of spontaneous speech would be essential to address these limitations and in turn ascertain their impact on individual performance.

The scope of exploration between acoustic data and animal group membership was also limited in the present set of systems, since it was only possible to examine *doves* and *worms* in any level of detail. As only three *phantoms* and no *chameleons* were identified across all tested systems, it was not possible to properly conduct an examination of the underlying data that gave rise to these classifications, though it is interesting to note that the *phantoms*, like some of the *doves*, were characterised by bimodal distributions. That only the *phantoms* performed relatively poorly in same-speaker comparisons suggests the possibility that bimodality was consistent between the questioned and known samples of those *dove* speakers, but not so in the case of the *phantoms*. Such possibility remains to be empirically tested in future studies. This study thus invites further research involving systems of other voice and speech features that focuses on the individual, in order to explore this issue more fully.

## 7. *Conclusion*

This exploratory study demonstrates the diagnostic value that individual-level analysis can add to LR-based system testing of (semi-automatic) linguistic-phonetic features,

providing insights into system performance that go beyond the level of single, global metrics. On the one hand, through identifying speakers with outlying performance within a system, it allows the potential factors behind their performance and the nature of errors to be investigated. On the other hand, it illustrates how speaker-discriminatory information from individual features combines in a complementary fashion. Looking at the trees, then, can be an essential way to enrich our understanding of the forest. The findings here support the recommendation by Dunstone, Yager (2009) and Alexander et al. (2014) that zooplot analysis be conducted in system evaluation and call for wider adoption of the practice.

## *Bibliography*

Alexander, A., Forth, O., Nash, J. & Yager, N. (2014). Zoo plots for speaker recognition with tall and fat animals. Paper presented at *23rd Annual Conference for the International Association for Forensic Phonetics and Acoustics*, August 31-September 3, 2014.

Becker, T., Jessen, M. & Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian mixture models. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, 22-26 September 2008, 1505-1508.

Boersma, P., Weenink, D. (2016). Praat: Doing phonetics by computer. [Computer programme] Version 6.0.19, retrieved 13 June 2016 from http://www.praat.org/.

Braun, A. (1995). Fundamental frequency: How speaker-specific is it? In Braun, A., Köster, J.-P. (Eds.). *Studies in forensic phonetics*. Trier: Wissenschaftlicher Verlag, 9-23.

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

Byrne, C., Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. In *International Journal of Speech, Language and the Law*, 11(1), 83-102. https://doi.org/10.1558/ijsll.v11i1.83

Cho, S., Munro, M.J. (2017). F0, long-term formants and LTAS in Korean-English bilinguals. In *Proceedings of the 31st General Meeting of the Phonetic Society of Japan*, Tokyo, Japan, 30 September-1 October 2017, 188-193.

Doddington, G., Liggett, W., Martin, A., Przybocki, M. & Reynolds, D. (1998). SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 Speaker Recognition Evaluation. In *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, 30 November-4 December 1998, Paper 0608.

Dunstone, T., Yager, N. (2009). *Biometric system and data analysis: Design, evaluation, and data mining*. New York: Springer.

Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd Ed.). New York: Harper & Row.

French, P., Foulkes, P., Harrison, P., Hughes, V., San Segundo, E. & Stevens, L. (2015). The vocal tract as a biometric: Output measures, interrelationships, and

efficacy. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK, 10-14 August 2015, Paper 0817.

Gold, E., French, P. & Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. In *Proceedings of Meetings on Acoustics*, 19, 060041. https://doi.org/10.1017/S0025100313000248

Hughes, H., Wood, S. & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. In *International Journal of Speech, Language and the Law*, 23(1), 99-132. https://doi.org/10.1558/ijsll.v23i1.29874

Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C. & San Segundo, E. (2018). The individual and the system: Assessing the stability of the output of a semi-automatic forensic voice comparison system. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, 2-6 September 2018, 227-231. https://doi.org/10.21437/Interspeech.2018-1649

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, 20-24 August 2017, 498-502. https://doi.org/10.21437/Interspeech.2017-1386

McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. In *International Journal of Speech, Language and the Law*, 11(1), 103-130. https://doi.org/10.1558/sll.2004.11.1.103

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. In *The Phonetician*, 101, 7-24.

Morrison, G.S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. In *Australian Journal of Forensic Sciences*, 45(2), 173-197. https://doi.org/10.1080/00450618.2012.733025

Morrison, G.S., Enzinger, E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_eval_01*) – Introduction. In *Speech Communication*, 85, 119-126. https://doi.org/10.1016/j.specom.2019.06.007

Nash, J. (2019). The effect of acoustic variability on automatic speaker recognition systems. PhD dissertation, University of York.

Nolan, F., Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. In *International Journal of Speech, Language and the Law*, 12(2), 143-173. https://doi.org/10.1558/sll.2005.12.2.143

O'Connor, K., Elliott, S., Sutton, M. & Dyrenfurth, M. (2015). Stability of individuals in a fingerprint system across force levels. In *Journal of Information Technology in Industry*, 3(2), 46-53.

R Core Team. (2018). R: A language and environment for statistical learning. Version 3.5.1.

Reynolds, D.A., Quatieri, T.F. & Dunn, R.B. (2000). Speaker verification using adapted Gaussian mixture models. In *Digital Signal Processing*, 10(1-3), 19-41. https://doi.org/10.1006/dspr.1999.0361

Rhodes, R. (2012). Assessing the strength of non-contemporaneous forensic speech evidence. PhD dissertation, University of York.

Rose, P. (2003). The technical comparison of forensic voice samples. In Freckelton, I., Selby, H. (Eds.). *Expert Evidence*, Sydney: Thompson Lawbook Co, Ch. 99.

Royal Canadian Mounted Police. (2010-2016). Voice ID Database [Unpublished audio corpus]. Collected at the University of Ottawa.

San Segundo, E., Tsanas, A. & Gómez-Vilda, P. (2017). Euclidean Distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics. In *Forensic Science International*, 270, 25-38. https://doi.org/10.1016/j.forsciint.2016.11.020

Scrucca, L., Fop, M., Murphy, T.B. & Raftery, A.E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. In *The R Journal*, 8(1), 289-317. https://doi.org/10.32614/RJ-2016-021

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