Beyond the Long-term Mean: Exploring the Potential of F0 Distribution Parameters in Traditional Forensic Speaker Recognition.

Yuko Kinoshita[1], Shunichi Ishihara[2], and Phil Rose[3]

[1]School of Languages and International Studies, University of Canberra, Australia
Yuko.Kinoshita@canberra.edu.au
[2]Faculty of Asian Studies, The Australian National University, Australia
shunichi.ishihara@anu.edu.au
[3]School of Language Studies, Faculty of Arts, The Australian National University, Australia
philip.rose@anu.edu.au

Abstract

Despite its many *prima facie* attractive properties for Forensic Speaker Recognition, F0 is regarded as having limited forensic value due to its large within-speaker variability. However, its forensic use to date has been limited mostly to its long-term mean and standard deviation. This paper examines the discriminatory potential, within a Likelihood Ratio-based approach, of additional parametric features from the distribution of long-term F0: its skew, kurtosis, modal F0 and modal density. Motivated by the observation that the overall long-term F0 distribution shows less within-speaker occasion-to-occasion difference, we report a forensic discrimination experiment with non-contemporaneous speech samples from 201 male Japanese speakers. Using a multivariate LR as discriminant distance with the six LTF0 distribution parameters, an EER of 10.7% is obtained from 201 target and 80400 non-target trials. We also investigate how the EER degrades as a function of amount of voiced speech.

1 Introduction

F0 is a popular parameter in traditional FSR. Its popularity probably stems from promising results in early SR research [1], together with its conforming to three of Nolan's desiderata for FSR parameters, namely: robustness, measurability, and availability [2]. The recordings that practitioners have to work with in actual forensic cases are often poor in quality and quantity. This can limit a comparison based on formant analysis – one of the most commonly used traditional parameter sets – severely. F0, however, is relatively robust against poor recording quality and differences in transmission channel. F0 is also an easier parameter to extract and measure than others, such as formants. Furthermore, in non-tone languages at least, F0 is not affected by the lexical content of the speech samples, so there is no need to locate comparable words or phonemes (although it is possible of course to compare samples with respect to F0 on sections with intonationally comparable structure). F0 is thus a very attractive parameter for traditional Forensic Speaker Recognition practitioners.

On the other hand, quite apart from the many linguistic uses of F0, which encodes tone, intonation and stress, many non-linguistic factors are known to affect it, including: state of health, emotional changes, discourse genre, noisiness of the environment, and whether or not the person is on the phone [3], [4], [5]. Thus many (eg. [6], [7]) have noted that a single speaker can show large variation in F0 from occasion to occasion, and even within a single recording session. Since the inherent strength of forensic speaker recognition parameters relies primarily on the ratio of within- to between-speaker variance, F0 is considered not very effective as a FSR parameter, and although some (for instance [2], [8], [9]), have suggested F0 as a potential speaker identification parameter, Kinoshita has demonstrated that because of its poor variance ratio, mean LTF0 shows very poor strength of evidence, typically generating Likelihood Ratios of effectively unity [10].

However, forensic analysis of F0 has concentrated so far on its long-term mean and standard deviation. If we could extract parameters which are less susceptible to within-speaker variation, of course, it could make a useful contribution to the field. This paper thus explores the forensic discriminatory potential of F0, focusing on the overall shape of its long-term distribution.

1.1 Distribution of long-term F0

As mentioned in the previous section, many non-linguistic and para-linguistic factors can strongly affect F0, and it thus must be considered to have limited use in FSR despite its popularity. However, observing the overall shapes of speakers' LTF0 distributions, we noted that they remained relatively consistent across different recording sessions. Figure 1 illustrates this with F0 distributions, each with a typical positive skew, from four speakers, sampled from each speaker in two separate recording sessions.

In the top two panels of figure 1, both speakers' two distributions show striking similarities in their shapes (while remaining different from each other). With these two particular speakers even mean LTF0 and SD seem to be quite stable across sessions, although in previous studies, such as [10], mean LTF0 was found to have a rather poor ratio of within- to between-speaker variation.

The bottom two panels of figure 1, however, are a different kettle of fish. The two distributions of the speaker in the bottom left panel are very similar in shape, but are shifted in frequency. If we are relying solely on long-term mean F0, two samples from this speaker probably would not have produced a strong LR, despite the similarity in the shape of the LTF0 distribution. This observation motivated us to include the simple statistical parameters which capture the shape of the F0 distribution.

Having said that, we also noticed that some speakers actually had large apparently occasion-dependent within-speaker variation in the shapes of their F0 distributions. The

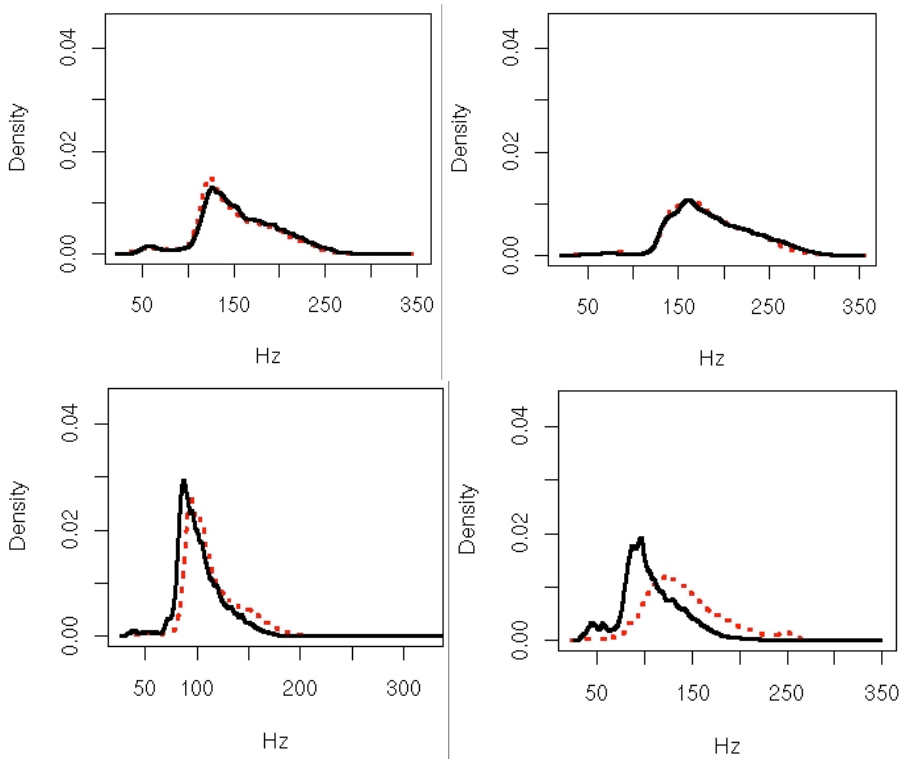


Figure 1: Four different speakers' LTF0 distributions, elicited on two separate occasions.

bottom right panel in figure 1 is an example. This implies that the shape of the distribution will not necessarily work for every case (but then again this is typical of the nature of variation of the human voice!). Considering the advantages of being able to use F0 in FSR, it would be worthwhile to investigate how useful the shape of a LTF0 distribution can be, using a relatively large forensically realistic dataset.

1.2 Database and Speaker Selection

For this study, we used male Japanese speakers selected from the Corpus of Spontaneous Japanese (CSJ) [11]. CSJ is a database which consists of various styles of speech recorded from 1464 speakers. The majority of the recordings was made in the style of either Academic Presentation Speech (APS) or Simulated Public Speech (SPS). APS was mainly recorded live at academic presentations, most of which were 12-25 minutes long. For SPS, 10-12 minutes mock speeches on everyday topics were recorded. In CSJ, all recordings were made using DAT and down-sampled to 16 kHz, with 16 bit accuracy. CSJ incorporates a five-scale evaluation of various aspects of the recordings. We used one of those evaluations, the so-called spontaneity scale, in order to select speakers. By *spontaneous*, CSJ means 'sounding as if it is not read out'. (In a situation such as an academic presentation, it is not uncommon for presenters to actually read out their prepared scripts.) Since FSR research requires us to work with forensically realistic data, we first selected speakers who were ranked highly (four or five) on the 1-5 spontaneity rating. The first two authors, who are native Japanese speakers, listened to a selection of the highly ranked recordings. They can confirm that they do indeed sound natural and not read out.

The other criterion for speaker selection was the availability of non-contemporaneous recordings. Our spontaneous sounding speakers had to have been recorded on two or more different occasions in order for us to attempt a forensically realistic discrimination. On the basis of these two criteria, then, we selected 201 male speakers, with two non-contemporaneous recordings for each speaker.

1.3 F0 Extraction and Parameterisation

F0 was extracted using the ESPS routine of the Snack Sound Toolkit [12] with Tcl at every 0.005 second. CSJ usefully annotates non-speech noise with a noise tag. The sections with this noise tag were excluded from the data.

The distributions of the extracted F0 were then parameterised. As well as long-term mean and SD, four other parameters, which relate to the shapes of the F0 distributions, were calculated for each of the 402 recordings. They are kurtosis, skew, modal F0, and the modal density.

Kurtosis measures peakedness of a distribution and skew measures its degree of symmetry. They are thus useful measures to characterise the overall shapes of the distributions. The mode will show what is the most commonly occurring F0 for each recording, and the density of it represents how concentrated it is.

In order to extract modal F0 and the modal density, firstly the probability density of the sampled F0 for each recording was estimated using binned kernel density (with the *bkde* function of R's KernSmooth library). The appropriate kernel density bandwidth was selected using direct plug-in methodology (the *dpik* function of R's KernSmooth library) [13, 14].

2 Likelihood Ratio-based approach

2.1 Likelihood Ratio

As is now well-known, the Likelihood Ratio (LR) is the probability that the evidence would occur if an assertion is true, relative to the probability that the evidence would occur if the assertion is not true [15]. In the context of forensic speaker recognition, it will be the probability of observing the difference between the suspect and offender speech samples if they had come from the same speaker (i.e. if the prosecution hypothesis were true) relative to the probability of observing the same evidence if it had been produced by different individuals (i.e. if the defence hypothesis were true). Letting *P* represent probability, *E* evidence, and *H* hypothesis, this can be expressed as (1):

$$LR = \frac{P(E \mid H)}{P(E \mid \overline{H})} \quad (1)$$

The LR will be larger than unity when the given evidence supports the hypothesis, and smaller than unity when the evidence does not support the hypothesis. The relative distance of the LR from unity quantifies the strength of the evidence.

It is also common practice to express the LR logarithmically, in which case the neutral value is of course 0, not unity. This seems easier for a layperson to grasp intuitively than a system in which, say, LRs of 10 and 0.1 have equal strength.

2.2 Likelihood Ratios in Forensic Science

In their introductory textbook on evaluating evidence, Robertson and Vignaux give two reasons why LRs should be used for evidence evaluation and presentation [15]. Firstly, the majority of evidence submitted to the Court is by nature only indicative, not determinative. The other reason is a result of the expert's role in the legal system. They are not in a position, either legally or logically, to make a decision on whether or not the defendant is guilty: this is the job of juries (or judges in some judicial systems). The expert will be violating ultimate issue rules if they do, and in any case they cannot, by Bayes' Theorem, estimate the probability of the hypothesis given the evidence, unless they know the prior odds in favour of the hypothesis, which they usually do not. The task of the FSR expert is thus to estimate the likelihood of observing the speech evidence when a particular hypothesis — usually the prosecution's — is correct versus when it is incorrect: that is, to estimate an LR for the speech evidence.

In addition to appropriateness for the legal system, LRs have another feature in evidence presentation: they allow evidence of different types to be combined. It is straightforward to combine multiple LRs from different evidence types by applying Bayes' Theorem, providing the evidence is not correlated. This is a significant feature, as most court cases involve many different types of evidence. This becomes even more significant in the evaluation of speaker identity: human speech is the product of such a highly complex system that no single parameter can distinguish one speaker from another consistently and reliably. It is thus essential to incorporate an adequate number of parameters in order to evaluate speech evidence (for instance, see [16]), and the use of LRs and Bayes' Theorem facilitates this.

Finally, of course - although Robertson and Vignaux did not foresee this - LRs are perfectly suited to testing how well same-speaker speech samples can be discriminated from different-speaker speech samples and they are now commonly used in forensically motivated discrimination experiments, both with automatic and traditional features [26]. This marks a very important development which has demonstrated the testability of the approach – one of the well-known *Daubert* criteria.

2.3 Multivariate Likelihood Ratios

The use of LRs in FSR with traditional features was first explored with univariate methods in "independence Bayes" fashion. Previous studies, such as [17], [18], [19], demonstrated that univariate LRs could indeed be used to discriminate same-speaker from different-speaker speech samples. One of the main problems with this approach was of course the possible correlation between predictor variables. In the aforementioned studies, some care was taken to avoid combining variables which were clearly correlated, but it is obviously of importance to be able to take correlation into account. This is, however, an unsatisfactory solution for two reasons. Firstly, we cannot assume that having no statistically significant correlation means the absence of correlation. Secondly, in actual forensic cases, experts often work with samples very limited both in quality and quantity. Under such circumstances, every parameter is precious. If the expert had to exclude some parameters altogether because of possible correlation, this could vitiate the whole enterprise.

The problem of estimating LRs from correlated variables was addressed by Aitken and Lucy by deriving multivariate LR (MVLR) formulae [20]. With MVLRs we can combine traditional parameters that may be strongly correlated, such as F2 and F3 in high front vowels. It is currently possible to model the reference population either as normal or with a kernel density. The MVLR still has problems such as its only accommodating two levels of variance, but it represents a significant step forward from the previous univariate formulae and has also been used in LR-based FSR discrimination experiments, e.g. [25].

The MVLR formula we used is taken from [20] and models the reference population with a Gaussian kernel density. Its numerator and denominator are given at (2) and (3).

3 Experiments

In this paper, we addressed two forensically motivated questions:

1) Can non-contemporaneous same-speaker speech samples be usefully discriminated from different-speaker speech samples on the basis of parameters derived from their long-term F0 distributions?
2) If so, how is the discrimination performance compromised by the amount of speech available?

In all experiments, two types of speaker pairs, non-contemporaneous same-speaker pairs and different-speaker pairs, were compared and evaluated using a MVLR as discriminant function. With 201 speakers we had 201 same-speaker comparisons (or target trials) and 80400 different-speaker comparisons (201 speakers produced 20100 combinations of speakers, and each different-speaker pair produced four different actual comparisons; i.e. Speaker A recording 1 *vs* Speaker B recording 1; Spk. A rec. 1 *vs* Spk. B rec. 2; Spk. A rec. 2 vs Spk. B rec. 1, Spk. A rec. 2 vs Spk. B rec. 2).

numerator of MVLR = (2)

$$(2\pi)^{-p}|D_1|^{-1/2}|D_2|^{-1/2}|C|^{-1/2}\left(mh^p\right)^{-1}\left|D_1^{-1}+D_2^{-1}+\left(h^2C\right)^{-1}\right|^{-1/2}$$
$$\times\exp\left\{-\frac{1}{2}\left(\bar{y}_1-\bar{y}_2\right)^{\mathrm{T}}\left(D_1+D_2\right)^{-1}\left(\bar{y}_1-\bar{y}_2\right)\right\}$$
$$\times\sum_{i=1}^{m}\exp\left[-\frac{1}{2}\left(y*-\bar{x}_i\right)^{\mathrm{T}}\left\{\left(D_1^{-1}+D_2^{-1}\right)^{-1}+\left(h^2C\right)\right\}^{-1}\left(y*-\bar{x}_i\right)\right]$$

denominator of MVLR = (3)

$$(2\pi)^{-p}|C|^{-1}\left(mh^p\right)^{-2}\prod_{l=1}^{2}\left[\left|D_l\right|^{-1/2}\left|D_l^{-1}+\left(h^2C\right)^{-1}\right|^{-1/2}\times\sum_{i=1}^{m}\exp\left\{-\frac{1}{2}\left(\bar{y}_l-\bar{x}_i\right)^{\mathrm{T}}\left(D_l+h^2C\right)^{-1}\left(\bar{y}_l-\bar{x}_i\right)\right\}\right]$$

where U, C = within-, between-speaker variance/covariance matrices; n_1, n_2 = number of replicates per speaker
m = number of speakers in reference population; p = number of assumed correlated variables per speaker
$D_l = D_1, D_2$ = offender, suspect var/cov matrices = $n_1^{-1}U, n_2^{-1}U$
h = optimal smoothing parameter for kernel density = $\left(4/(2p+1)\right)^{1/(p+4)}m^{-1/(p+4)}$
$\bar{y}_l = \bar{y}_1, \bar{y}_2$ = offender, suspect means; y^* = $\left(D_1^{-1}+D_2^{-1}\right)^{-1}\left(D_1^{-1}\bar{y}_1+D_2^{-1}\bar{y}_2\right)$
$\bar{x}_i$ = within-speaker means of reference population.

The discrimination test was performed intrinsically, assuming that the effect of the test data not being independent from the population data would be negligible in this research, because of the large number of speakers involved. We therefore did not use a 'leave one out' approach.

3.1 Experiment 1

This experiment examined the overall usefulness of the LTF0 distribution parameter set. We incorporated all six parameters using the MVLR formula, and used all available speech. This means that, in this experiment, each recording had a different length of speech but mostly they lay between 10 and 25 minutes. Table 1 summarises the results of the testing on the 201 same-speaker pairs and 80400 different-speaker pairs. "S-S" in the first column indicates same speaker comparisons, and "D-S" shows different speaker comparisons. The column labelled "Correct" indicates the number of the comparisons which supported the hypothesis that was consistent with the reality.

Table 1: Summary of MVLR-based discrimination using six LTF0 parameters.

	Total	Correct	%	EER
S-S	201	181	90.0	10.7
D-S	80400	72456	90.1	

As can be seen, results are promising. Using six features, we obtained an EER of 10.7%: approximately 72,000 out of 80,400 (90.1%) different-speaker trials and 181 out of 201 (90.0%) same-speaker trials produced MVLRs supporting a hypothesis consistent with reality. This is a considerably better performance in the different-speaker comparisons than the previously reported results. Kinoshita conducted a similar experiment using long-term F0 of 12 male Japanese speakers and obtained similar results for the same-speaker trials (90% supporting the hypothesis that is true in reality). The different-speaker trials, however, were abysmal, with only 32.6% that supported the hypothesis consistent with the actual speakers' identity [10].

As is now conventional with forensic speaker discrimination, the results are also presented as a Tippett, or reliability, plot in figure 2. Same-speaker comparisons are shown by the dashed line, different-speaker comparisons by the solid lines (the four curves show the four different comparisons made within each different-speaker combination). The horizontal axis shows logLR greater than … , and the vertical axis shows number of trials (with same-speaker trials plotted inversely). Thus it can be seen for example that about 10% of same-speaker trials were incorrectly evaluated with LRs more likely had they come from different speakers.

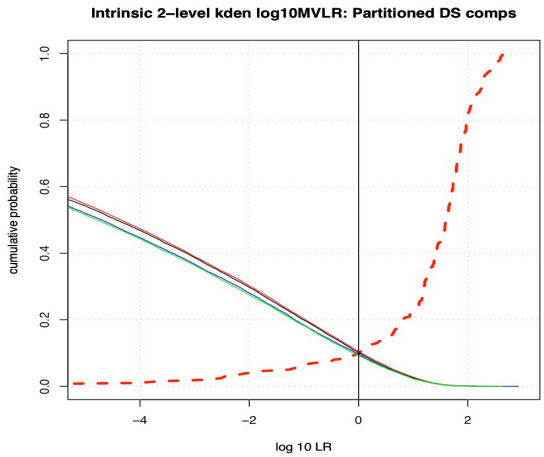


Figure 2: Tippett plot for MVLR-based discrimination using six LTF0 parameters

The Tippett plot in figure 2 presents several characteristics typical for MVLR discrimination using traditional features. Firstly the range of correctly evaluated same-speaker samples is very much smaller than for different-speaker samples. The most favourable same-speaker pairs were only a little more

than 400 times more likely assuming same-speaker provenance, whereas the best different-speaker trials have astronomically low LRs. The location of the EER very near to logLR= 0 is another common feature, and probably reflects the derivational nature of the MVLR formula.

The rigorous information-theoretic evaluation of overall LR-based discriminant performance has recently received useful and welcome attention [24]. However, although we intend to look at the calibration- and discrimination-loss decomposition of our results, with associated APE plots, in this paper we feel it is more useful to focus locally, on two types of comparisons: the combination producing the best performance (ie. strongest LR supporting the correct hypothesis) and those supporting the counterfactual hypothesis. First of all, the best performances. The best performance here means the maximum logLR for same-speaker comparisons and minimum logLR for the different-speaker comparisons. These are given in Table 2.

Table 2: LogLR values for the best comparisons

	S-S	**D-S**
Speaker ID	280	65 *vs* 173 (combination 1)
Log LR	2.648	-76.89

The three speakers' F0 distributions involved are presented in figures 3 and 4.

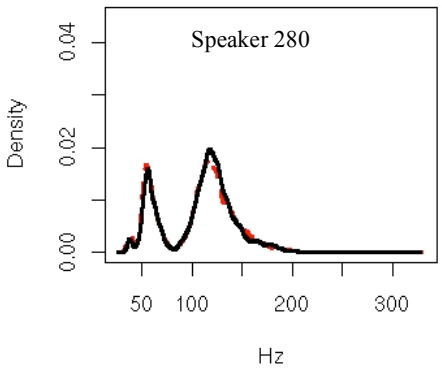


Figure 3: Best performing same-speaker comparison (logLR = 2.65)

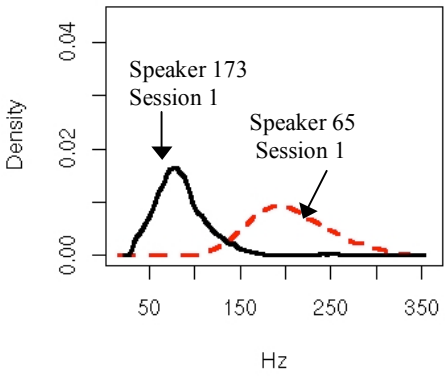


Figure 4: Best performing different-speaker comparison (LogLR= -77.0)

As mentioned, comparing the absolute values of these two logLRs, it is apparent that the same-speaker comparisons produce much weaker logLRs than the different-speaker comparisons. Champod and Evett have proposed a verbal scale to assist interpretation of LRs [21]. In their scale, logLR 2 to 3 (or -3 to -2) indicates "moderately strong" evidence, and logLR 3 to 4 (or -4 to -3) indicates that the evidence is "strong". The value 2.65 thus counts as only "moderately strong", whereas, -76.89 is of course an extremely strong value and basically off the chart. Considering the nature of within- and between-speaker variation, however, this discrepancy is not such a surprising result. Two samples cannot be any more similar than being the 'same' (and under these circumstances the magnitude of the LR is determined by other MVLR terms like the variance ratio and the number of items in the samples). On the other hand, for the variation between different samples, both intrinsic anatomical factors (in the case of F0, the length and mass of speakers' cords) and the extrinsic use to which they are put (to produce habitually creaky phonation, for example) contribute to the degree of difference. This means different speakers can differ from each other to an effectively limitless extent. However, it is important to note that we would not get such strong values for different-speaker comparisons in reality anyway, because two voices would probably not be compared in the first place if they sounded as different as implied by these large negative LRs. In the case under consideration, for example, the two speakers' mean F0 values differed by over 100 Hz (205.2 Hz for Speaker 65, 87.5 Hz for Speaker 173). Assuming that the recordings were made under comparable conditions, two samples showing such a great difference would sound so very different that they would be unlikely to arouse suspicion in a real case.

Now we turn to the erroneous comparisons. As mentioned earlier, the LR is a continuous expression of how much more likely it is to observe the given evidence when the hypothesis is true than when it is not. The LR thus by nature incorporates the possibility of not supporting a hypothesis which is true in reality: that simply means that the difference between the speech samples is greater (or lesser) than would be expected on the basis of the between- and within-speaker variance of the reference population. Table 3 summarises the comparisons which strongly supported counterfactual hypotheses and the size of the logLR associated with them.

Table 3: Erroneous classifications and associated logLR magnitude.

	S-S			**D-S**		
range	**< 0**	**< -2**	**< -3**	**> 0**	**> 2**	**> 3**
no	20	9	4	7944	56	0
%	9.95	4.48	1.99	9.88	0.07	0

We cannot conclude that an LR is wrong on the grounds that it supported the hypothesis which turned out to be wrong – LRs are part of Bayes' Theorem and a theorem cannot be 'wrong'. However, if a statistical model produced a great number of very strong LR estimates which support a wrong hypothesis, we would be inclined to question either the validity of the model or the inherent discriminability of the medium, or both. Realistically, we do not know what the expected range of logLR is when it is supporting the wrong hypothesis. However, referring to Champod and Evett's scale, we decided to look into the cases where absolute values of logLR are above 3 (i.e. "moderately strong").

In this experiment, different-speaker comparisons did not produce as extreme counterfactual logLRs as the aforementioned comparison between Speakers 65 and 173.

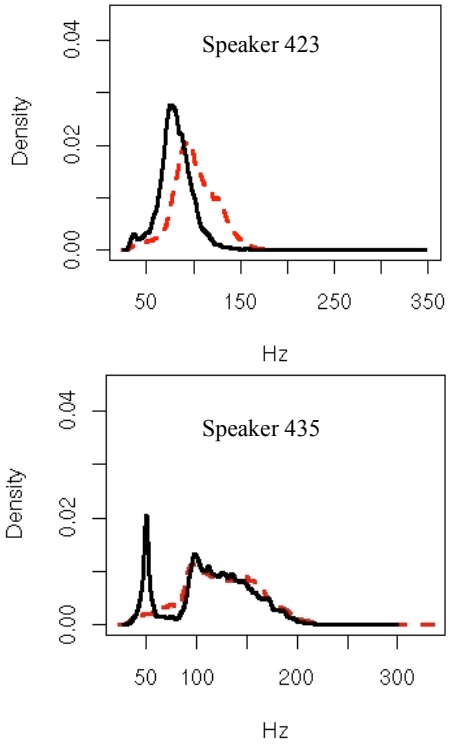
Although about 10% of the different-speaker comparisons produced logLRs greater than 0, they are all below three. In other words, there is no instance of supporting the wrong hypothesis strongly. This relatively low incidence of false positives is of course a desirable property in the forensic speaker recognition context.

The same-speaker comparisons, on the other hand, have a few instances of worrisome results. Table 4 gives the four same-speaker comparisons which produced logLRs smaller than -3.

Table 4: Summary of logLR smaller than -3.

Speaker ID	423	435	1180	397
logLR	-9.011	-3.665	-4.046	-6.95

The F0 distributions from those four poorly evaluated speakers are also given in Figure 5. At first glance, perhaps with the exception of speaker 423, the two distributions for each speaker do not look so different. However, the very small logLR for speaker 435 probably can be explained by the very strong peak at a very low frequency region. This was presumably to do with his audibly heavy creakiness on one occasion but not in the other. This has of course made the distribution of his two recordings very different in their mean, modal F0 and probability density of the modal F0. We need therefore to be able to ignore the lower peak when calculating some parameters. With respect to speakers 1180 and 1397, although both their recordings show a similar overall shape for their distribution, they do differ in probability density of the modal F0 (as indeed do the other two bad speakers). If it is the case that the difference in modal F0 probability density is responsible, then we need to investigate the relative discriminatory power of the individual distribution parameters with univariate LRs – a task for future research.



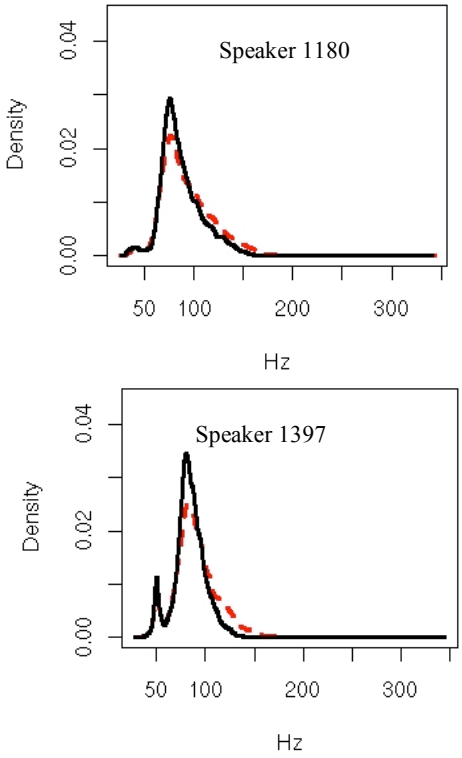


Figure 5: LTF0 distributions of problematic same-speaker comparisons

3.2 Experiment 2: Degradation function

In the second experiment, we investigated how the duration of the recordings affects the discrimination performance. This is an important consideration for FSR practitioners, since the samples available in actual forensic cases are often very short. It is thus useful to know how the MVLR-based discrimination performance degrades with decreasing amounts of available speech. This will give us an idea of how long forensic speech samples should be for F0 to be useful in FSR, in the sense of contributing a potentially useful strength of evidence.

Nolan notes that we need around one minute of speech to reliably capture individuals' within-speaker variation in F0 [2]. However, as Rose points out, this necessary duration would differ from language to language [22]. F0 can only be extracted from voiced segments, and the phonological structure of the language (its phonemic inventory, lexical incidence and phonotactics) determines the ratio between voiced and voiceless segments in an utterance. Catford shows that the voiced-to-voiceless ratio differs significantly from language to language [23]. For instance, in French, voiced segments constitute 78% of utterances, whereas only 41% of segments in Cantonese are voiced. Although Japanese does not appear in this list, we found from our 201 speakers that on average 70.0% (SD 4.48) of Japanese utterances are voiced. This is a similar ratio to that of English (72%) in Catford's list.

In this experiment, we again calculated LogMVLRs with the six parameters used in the previous experiment, but this time we incrementally reduced the amount of voiced speech for analysis.

Thirty-one different samples of voiced speech, ranging from five to 180 seconds, were generated from the beginning of each of the 402 recordings. The duration of the voiced speech was estimated as overall duration of speech * 0.7 according to the Japanese voicing ratio of 70% estimated above. The amounts of voiced speech were incrementally

reduced as follows. From 180 to 30 seconds, the duration was decreased by 15 seconds; and from 30 to 5 seconds by 1 second steps.

The EER as a function of the duration of voiced speech is presented in figure 6. It is clear firstly, but unsurprisingly, that performance improves with amount of voiced speech available. The figure seems to suggest that we can divide the improvement of the performance into three stages. From five to 30 seconds of voiced speech available, it seems that the EER improves steadily, from ca. 23% to 16%. With ca. 30 to 90 seconds of voiced speech, we again observe some improvement in EER, but the rate is slower than over the first 30 seconds. There is also the intriguing local hiccough at ca. 40 seconds of voiced speech. After about 90 seconds of voiced speech, the EER seems to have become asymptotic to ca. 12%, which is very close to our EER of 10.7% obtained for all available data.

In order to extract 90 seconds of voiced Japanese speech, we needed roughly 130 seconds of net speech. Thus the results of this experiment seem to suggest that for an optimum evaluation of Japanese data, we need samples of just over two minutes.

The most important results of this experiment, however, are firstly that it still appears possible to obtain a EER of between ca. 20% and 23% with a relatively small amount – less than 15 seconds – of voiced speech. Of course it will be necessary to examine the associated Tippett plots for these smaller amounts of speech, in order to see what sorts of LR ranges, and quality of calibration, are involved. Also it will be interesting to see for what amount of speech the EER becomes useless. Secondly, for small amounts of voiced speech (< 30 sec.) a small increase in the amount can make a relatively large difference in discrimination performance. It is therefore worth while in real cases to use as much voiced speech as possible.

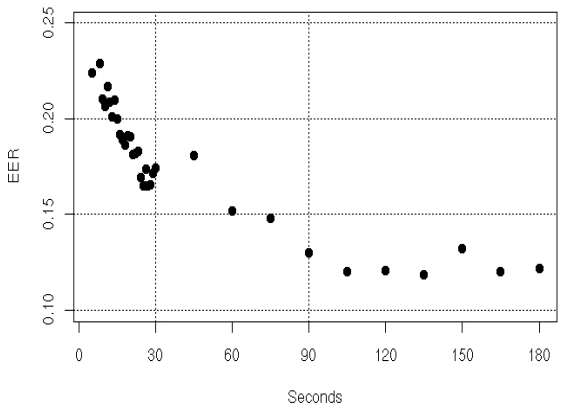


Figure 6: MVLR EER as a function of the duration of available voiced speech.

4 Summary and Way Ahead

In this study, we addressed the question of whether traditional forensic LR-based discrimination is possible using additional parameters from speakers' LTF0 distributions, in order to capture their gestalt similarity. The results were promising. By combining six parameters: mean and SD F0, kurtosis, skew, modal F0, and the probability density of the modal F0, we demonstrated an EER of 10.7% over a reasonably large set of target and non-target trials. This represented a significant improvement on methods relying solely on mean and SD F0. A consideration of the erroneous classifications suggested that we could refine the parameterisation. This suggests potential for further improvement in FSR discrimination within the LR-based testing paradigm.

An experiment on the relationship between EER and the amount of voiced speech revealed that, with Japanese, we need roughly about two minutes of speech to have near optimum FSR performance using F0, but that small amounts of voiced speech could still deliver EERs that might be associated with usable strengths of evidence. It was pointed out that this value could differ language-to-language, and further research is needed in other languages.

We are aware that the speech data used, although natural sounding, may still differ from actual forensic samples in ways that favour discriminability - it is after all monological and not conversational, and may because of this incorporate less within-speaker variation than conversational speech (although that has yet to be determined). This caveat should be born in mind.

We believe that, with some refinement of the methodology, its performance can further improve. Firstly, the treatment of creaky phonation needs more consideration. We set the F0 sampling range from 30Hz to 350Hz, hoping that the creakiness often observed in utterance-final position might contribute usefully to the speaker's LTF0 profile. In the calculation of long-term mean, we averaged everything sampled, ignoring the secondary peak which was present for some creaky speakers at a lower frequency range. This would no doubt have obscured the speaker profile in this parameter. As a next step, it would be useful to separate any secondary peak from the main part of the distribution and retest. Also, finding the most suitable method to parameterise this creak-related secondary peak and incorporate it in the analysis could improve the performance.

A final caveat is in order. Although this paper has focused on the *discriminatory* potential of the LTF0 parameter set, it should not be forgotten that the primary, forensic, use of the MVLR is not discriminatory, but as a means of estimating the strength of evidence in a specific case. The contrast between these approaches is clear if we consider LR values close to 0. In discrimination, a LR value either side of 0 counts as either correct or not, and contributes, in its small way, to the EER. But from the point of view of estimating the strength of evidence, such low values would mean the evidence is useless.

5 References

[1] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contour," *JASA,* vol. 52, pp. 1687-1697, 1972.

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

[3] K. Maekawa, "Phonetic and phonological characteristics of paralinguistic information in spoken Japanese," i*n Proc. of the 5th International Conference on Spoken Language Processing*, Sydney, 1998, paper no. 997.

[4] T. Watanabe, "Japanese pitch and mood," *Nihongakuho, Osaka University,* vol. 17, pp. 97-110, 1998.

[5] J. Elliott, "Comparing the acoustic properties of normal and shouted speech: a study in forensic phonetics," in *Proc. Eighth Australian International Conference on Speech Science ad Technology*, Canberra, 2000, pp. 154-159.

[6] P. French, "An overview of forensic phonetics with particular reference to speaker identification," *Forensic Linguistics,* pp. 169-181, 1994.

[7] A. Braun, "Fundamental Frequency - How Speaker Specific Is It?," *BEIPHOL Studies in Forensic Phonetics,* pp. 9-23, 1995.

[8] M. R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-23, pp. 178-182, 1975.

[9] M. Jiang, "Fundamental frequency vector for a speaker identification system," *Forensic Linguistics,* vol. 3, pp. 95-106, 1996.

[10] Y. Kinoshita, "Does Lindley's lr estimation formula work for speech data?: investigation using long-term f0," *International Journal of Speech Language and the Law,* vol. 12, pp. 235-254, 2005.

[11] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc.* of *the Second International Conference of Language Resources and Evaluation (LREC2000)*, Athens, 2000, pp. 947-952.

[12] K. Sjölander, "The Snack Sound Toolkit ", 2006.

[13] M. P. Wand and M. C. Jones, *Kernel Smoothing*. London: Chapman and Hall, 1995.

[14] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society,* vol. 53, pp. 683–690, 1991.

[15] B. Robertson and G. A. Vignaux, *Interpreting Evidence*. Chichester: Wiley, 1995.

[16] Y. Kinoshita, "How small can it get? Forensic speaker identification as a function of parameter number," in *Proc. Ninth Australian International Conference on Speech Science and Technology*, Melbourne, 2002.

[17] Y. Kinoshita, *Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach using Formants*, Ph.D. Thesis, Australian National University, 2001.

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

[19] T. Alderman, *Forensic Speaker Identification: A Likelihood Ratio-based Approach Using Vowel Formants,* LINCOM Studies in Phonetics 01, Munich: Lincom, 2005.

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

[21] C. Champod and D. Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication,* vol. 31, pp. 193-203, 2000.

[22] P. J. Rose, "How effective are long-term mean and standard deviation as normalisation parameters for tonal fundamental frequency?" *Speech Communication,* vol. 10, pp. 229-247, 1991.

[23] J. C. Catford, *Fundamental problems in phonetics*. Edinburgh: Edinburgh University Press, 1977.

[24] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language* Special Issue vol. 20, 2-3 pp. 230-275, 2006

[25] P. Rose, Y. Kinoshita and T. Alderman, "Realistic Extrinsic Forensic Speaker Discrimination with the Diphthong /ai/", in Warren P. & Watson C. Editors, *Proc. 11th Australian International Conference on Speech Science and Technology*, pp. 329 – 334, 2006.

[26] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano and J. Ortega-Garcia, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", *IEEE Transactions on Audio Speech and Language Processing*, vol. 15,7, pp. 2104-2115, 2007.

6 Acknowledgement

We thank our reviewers for their time and comments.