# The Effect of Sampling Procedures on the Performance of Likelihood Ratio Based Forensic Voice Comparison: F0 Distributional Parameters

*Shunichi Ishihara*
Department of Linguistics, the Australian National University
shunichi.ishihara@anu.edu.au

## Abstract

In this study, various sampling procedures were tested for F0 distributional parameters in order to see how the performance of the forensic voice comparison system is influenced. This was done by changing the width of analysis window and the degree of its shifting, resulting in different sets of feature vectors. The results show that the discriminability is fairly comparable across the different sampling procedures, whereas there is a large difference in calibration loss between the procedures if the analysis windows are not overlapped. It is reported that the use of overlapped analysis windows contributes to the stability in calibration loss.
**Index Terms**: forensic voice comparison, long-term F0 distribution, sampling procedures

## 1. Introduction

The usefulness and efficacy of the features based on long-term F0 distributional patterns (e.g. the mean, sd, skewness, kurtosis of the distribution) have been reported and demonstrated for forensic voice comparison (FVC) [1]. However, since the long-term F0 distribution is usually obtained from the entire speech sample available for the caseworker, only one set of feature vectors is consequently available for modelling. It goes without saying that the more sets, the better for accurately estimating the variance of a speaker [2]. Thus, a question naturally arises: isn't it better to obtain multiple numbers of feature vectors for each speaker by breaking up the recording into some chunks? Needless to say, there is a trade-off here between the amount of data for building the distributional model of each chunk and the number of the set of feature vectors. This study attempts to answer the above question, with the Multivariate Kernel Density (MVKD) likelihood ratio formula [3], which is one of the common procedures for FVC.

## 2. Likelihood ratio

The current study is a likelihood ratio (LR) based FVC study. For FVC, as expressed in Equation (1), LR is the probability of observing the difference (referred to as the evidence, $E$) between the offender's and the suspect's speech samples if they had come from the same speaker ($H_p$ = the prosecution hypothesis, is true) relative to the probability of observing the same evidence (E) if they had been produced by different speakers ($H_d$ = defence hypothesis, is true) [4].

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

The LR expresses the relative strength of the given evidence regarding the competing hypotheses ($H_p$ vs. $H_d$). It is a common practice to present LRs in the logarithmic scale (base 10), in which case the neutral point is the $\log_{10}LR = 0$.

## 3. Database and F0 extraction

The monologues stored in the Corpus of Spontaneous Japanese (CSJ) [5] are used in this study. The criteria for selecting speakers from the CSJ, the pre-processing (e.g. downsampling) of the selected speech samples and so on, are explained in detail in [1]. The selection criteria resulted in the selection of 201 speakers (201 speakers * 2 non-contemporaneous sessions = 402 speech samples), and they were divided into three mutually exclusive databases of test, background and development (each of which consists of 67 speakers).

The stream of speech in a recording is pre-annotated and chunked into the unit of *utterance* in the CSJ. Utterances are separated by silences with durations of 0.4 sec. or longer. The CSJ also annotates non-speech noise, and the sections marked with a noise tag were excluded from the F0 extraction. The ESPS routine of the Snack Sound Toolkit (http://www.speech.kth.se/snack/) was used to extract F0 at every 0.005 second from the utterances for each recording. The distributional pattern of the extracted F0 values was parameterised by calculating the following six features: the mean, standard deviation, skew, kurtosis, modal F0 and the density of the modal F0. The `KernSmooth` library of the `R` statistical package was used for the modal F0 and its density.

## 4. Sample sizes and sampling procedures

As can be seen in Table 1, eight different sample sizes are used in this study. They are given in terms of the numbers of F0 samples and their equivalent durations.

Table 1: *F0 sample sizes. Durations in sec.*

| Numbers | Durations |
|---|---|
| 2000 | 10 |
| 4000 | 20 |
| 6000 | 30 |
| 8000 | 40 |
| 12000 | 60 |
| 16000 | 80 |
| 20000 | 100 |
| 24000 | 120 |

For example, 2000 means that the first 2000 F0 values of a given recording are used to build a long-term F0 distribution, from which the six features are extracted. Since the F0 values were calculated at every 0.005 sec., the duration of the sample is equivalent to 10 sec. (= 2000 * 0.005). In order to avoid unnecessary confusion, the different samples sizes are referred to with their durations in sec.

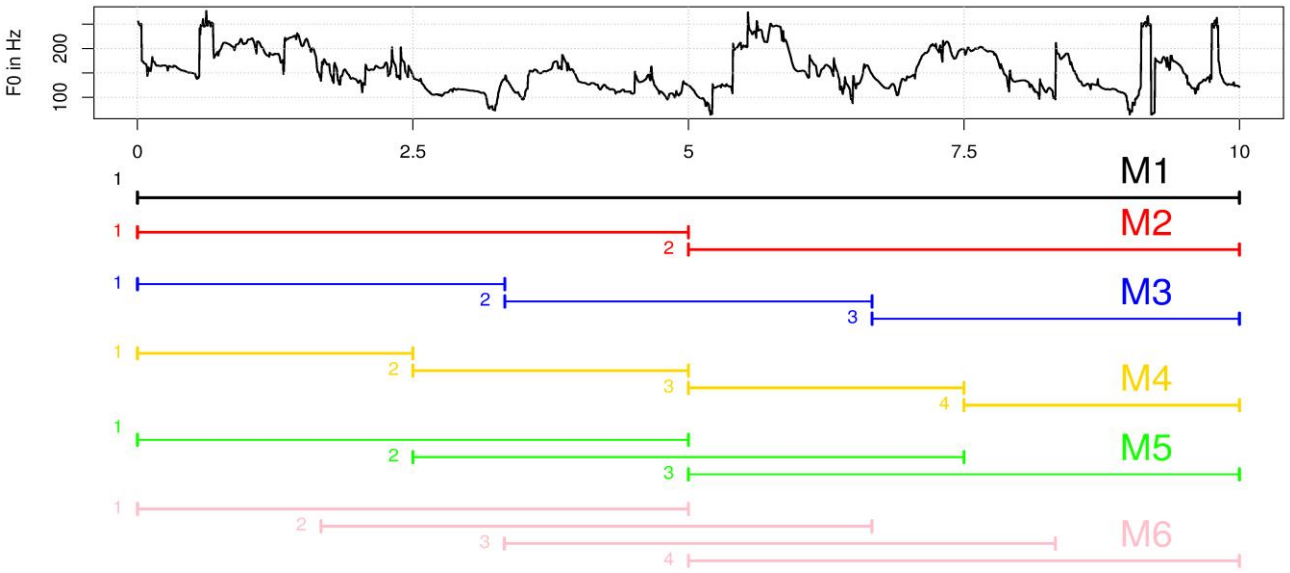


Figure 1: *Six sampling procedures (M1-M6) are schematically presented with sequentially plotted F0 values (x-axis = sec.).*

Six different sampling procedures are tried in this study. They (M1-M6) are schematically given in Figure 1. In the top half of Figure 1, the first 2000 F0 values extracted from an example recording are sequentially plotted (10 sec. in total). As can be seen in Figure 1, the six different sampling procedures have different analysis window sizes and different degrees of shifting (or overlaps), resulting in different amounts of feature vector sets. The F0 values which fall within each analysis window are pooled together in order to build the long-term F0 distribution. In M1-M4, for example, i) the size of each analysis window is different and ii) there is no overlap between the adjacent windows. In M1 (the benchmark case), the maximum size of analysis window is used; that is, only one set of feature vector. In M2, M3 and M4, the 50%, 33.3% and 25% of the maximum window size are used, resulting in two, three and four sets of feature vectors, respectively. In M5 and M6, half of the maximum window size is used with the analysis window being shifted by 50% and 33.3%; three and four sets of feature vectors, respectively.

## 5. Estimation of strength of evidence and performance assessment

The Multivariate Kernel Density (MVKD) formula [3] was used with a logistic-regression calibration [6] to estimate the likelihood ratios (LRs). In the MVKD formula, the covariance matrices ($D_l$, $l$ = 1,2) of the offender and the suspect samples are assumed to be constant, and they are estimated from the pooled within-speaker covariance matrix ($U$) of the background database, being scaled by the number of samples (= feature vectors) ($n$) ($n_l{}^{-1}U$, $l$ = 1,2). That is, only the suspect and offender means are used in the calculation of the LRs in the MVKD formula. Thus, when the difference between the feature vectors is compared by means of a *Mahalanobis* distance, the number of feature ($n$) vectors plays an important role.

A logistic-regression calibration is applied to the outcomes (customarily called *scores*) of the MVLR formula. The logistic-regression weight is obtained from the development database. For the different sample sizes, the performances of the six different sampling procedures given in Figure 1 are assessed by means of the log-likelihood-ratio-cost ($C_{llr}$), including $C_{llr}\_min$ and $C_{llr}\_cal$. The $C_{llr}$ measures the overall accuracy of an FVC system. The $C_{llr}\_min$ and $C_{llr}\_cal$ specifically examine the discrimination and calibration performances of the system, respectively [6]. The FoCal toolkit (https://sites.google.com/site/nikobrummer/focal) was used for both the calibration and $C_{llr}$.

## 6. Results and discussion

The $C_{llr}$, $C_{llr}\_min$ and $C_{llr}\_cal$ values are given in Figure 2 for each of the sampling procedures (M1-M6). The left-hand side of Figure 2 (Panels a, b and c) are for M1-M4, and the right-hand side (Panels d, e and f) are for M1, M5 and M6. M1 is the benchmark experiment. The different sizes of analysis window were used in M1, M2, M3 and M4 with 100% shift (= no overlapping between the windows). Half of the maximum window size was used in both M5 and M6 with different degrees of shifting (50% and 33.3%, respectively).

Judging from Figure 2a, in which the $C_{llr}$ values of M1-M4 are given, it is not that one procedure consistently performs better than the others; yet M2 (red) performed better than M1 (black) for all sample sizes, except the sample size of 80 sec. The performances of M1-M4 are relatively comparable up to the sample size of 40 sec., after which they largely fluctuate even within the same sampling procedures, resulting in that the best- or worst-performing sampling procedure is not consistent; for example, M1 achieved the best $C_{llr}$ (= 0.613) for the sample size of 80 sec. while M1 performed worst ($C_{llr}$ = 0.775) with the sample size of 100 sec. As for the sample sizes of 100 and 120 sec., M2-M4 (multiple sets of feature vectors) consistently performed better than M1 (one set of feature vector). An advantage of having multiple sets of feature vectors may start emerging when the sample size is relatively large (100 sec. or longer); perhaps each chunk is large enough to accurately estimate the distributional pattern. This point needs to be confirmed with a larger amount of data.

The observation of Figure 2bc, in which $C_{llr}\_min$ and $C_{llr}\_cal$ values are plotted, respectively, as a function of the sample size, tells that i) the discriminability of the system (Figure 2b) is fairly comparable across M1-M4, ii) the discriminability of the system (Figure 2b) generally improves

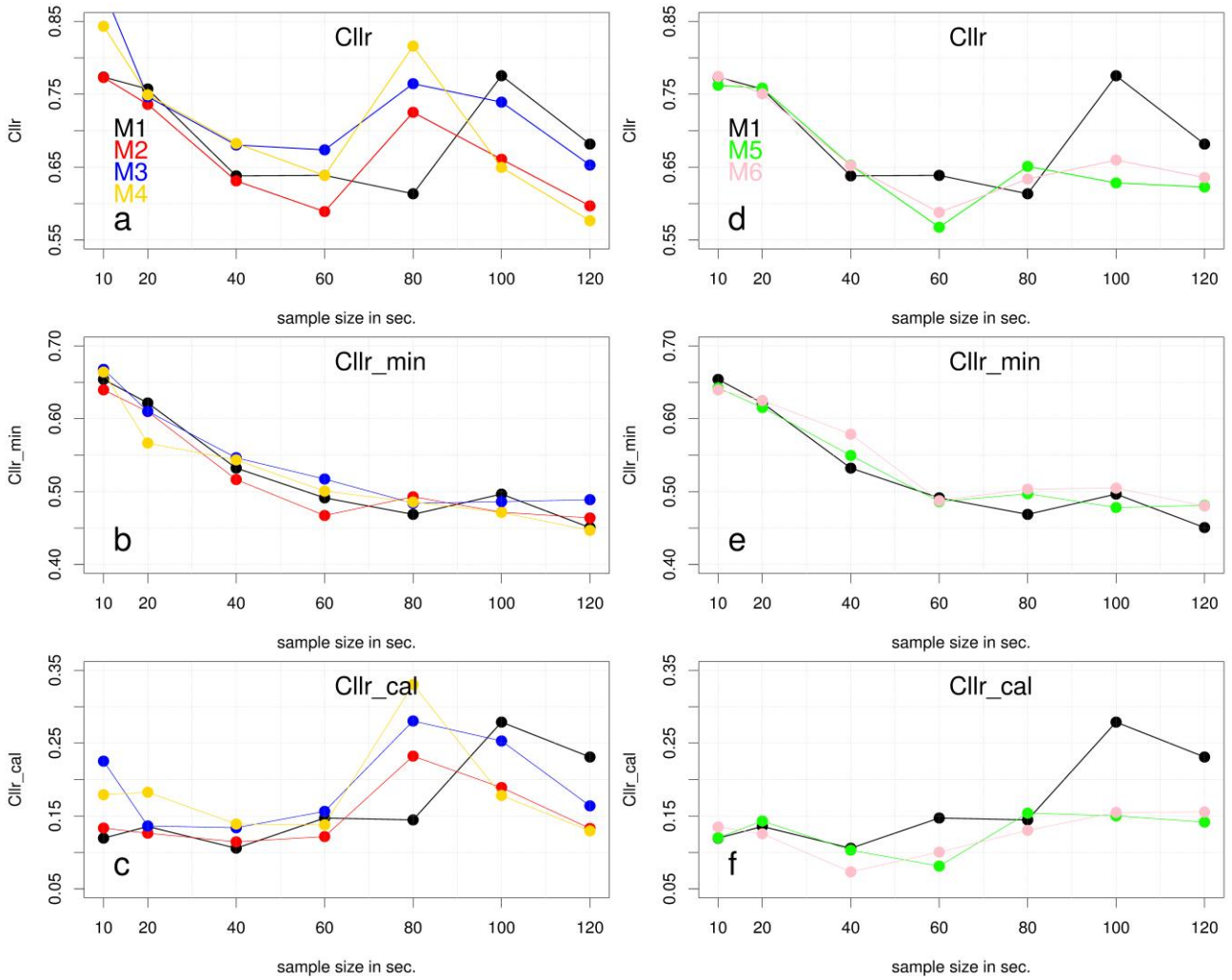


Figure 2: *The $C_{llr}$ (Panels a and d), $C_{llr}$_min (b and e) and $C_{llr}$_cal (c and f) values of the different sampling procedures (M1-M6) are plotted against the different sample sizes (10, 20, 40, 60, 80, 100 and 120 sec.). Panels a, b and c are for M1-M4, and Panels d, e and f are for M1, M5 and M6. The M1 is the benchmark experiment.*

as the amount of sample size increases until the sample size of 60 sec., after which the performance starts converging; this is something we naturally expect and iii) the calibration loss (Figure 2c) is not stable (the sample size of 80 sec. or longer).

It is interesting to know that the large fluctuations in performance in terms of $C_{llr}$, which is demonstrated in Figure 2a, in particular when the sample size is 60 sec. or longer, are due to large fluctuations in calibration performance. Although the reason behind this is not clear at this stage, it may be due to an inherited nature of the MVKD formula [7].

The results of M5 and M6 are plotted in Figure 2def together with the results of M1. M5 and M6 are different from M1-M4 in that the analysis window shifted in the way that adjacent windows overlap; the degree of overlap is 50% and 66.6%, respectively. Needless to say, overlapped analysis windows have a smoothing effect in sampling.

It is clear from Figure 2d that the performances of M5 and M6 are very similar in terms of $C_{llr}$, and also that the performance improves more or less in an expected manner as the sample size increases. This observation of M5 and M6 is quite different from that of M1-M4, which is not as stable. Furthermore, like the observation made for M1-M4 in Figure 2a, M5 and M6 performed better than M1 when the sample size is large (e.g. 100 and 120 sec.).

As for the discriminability potential of M5 and M6 (refer to Figure 2e), it shows the same trend as that of M1-M4. Unlike M1-M4, the calibration loss of M5 and M6 is relatively stable, as shown in Figure 2f, in that the $C_{llr}$_cal values sway around the narrow range between $C_{llr}$_cal = 0.05 and 0.15. It is clear that overlapped analysis windows contribute to the stability of the system performance.

Figure 3 contains the Tippett plots of M1, M2, M5 and M6 for the sample size of 120 sec. In terms of $C_{llr}$ values, M2 (0.576), M5 (0.622) and M6 (0.635) are fairly similar in performance, but they are better in performance than M1 (0.681). It is evident from Figure 3 that some large counter-factual same speaker LRs, which are indicated by the circle in Figure 3a, largely influence the $C_{llr}$ value; otherwise the magnitude of the derived LRs, including both consistent-with-fact and contrary-to-fact LRs, are fairly similar across the different sampling procedures. This observation is generally true for the other cases.

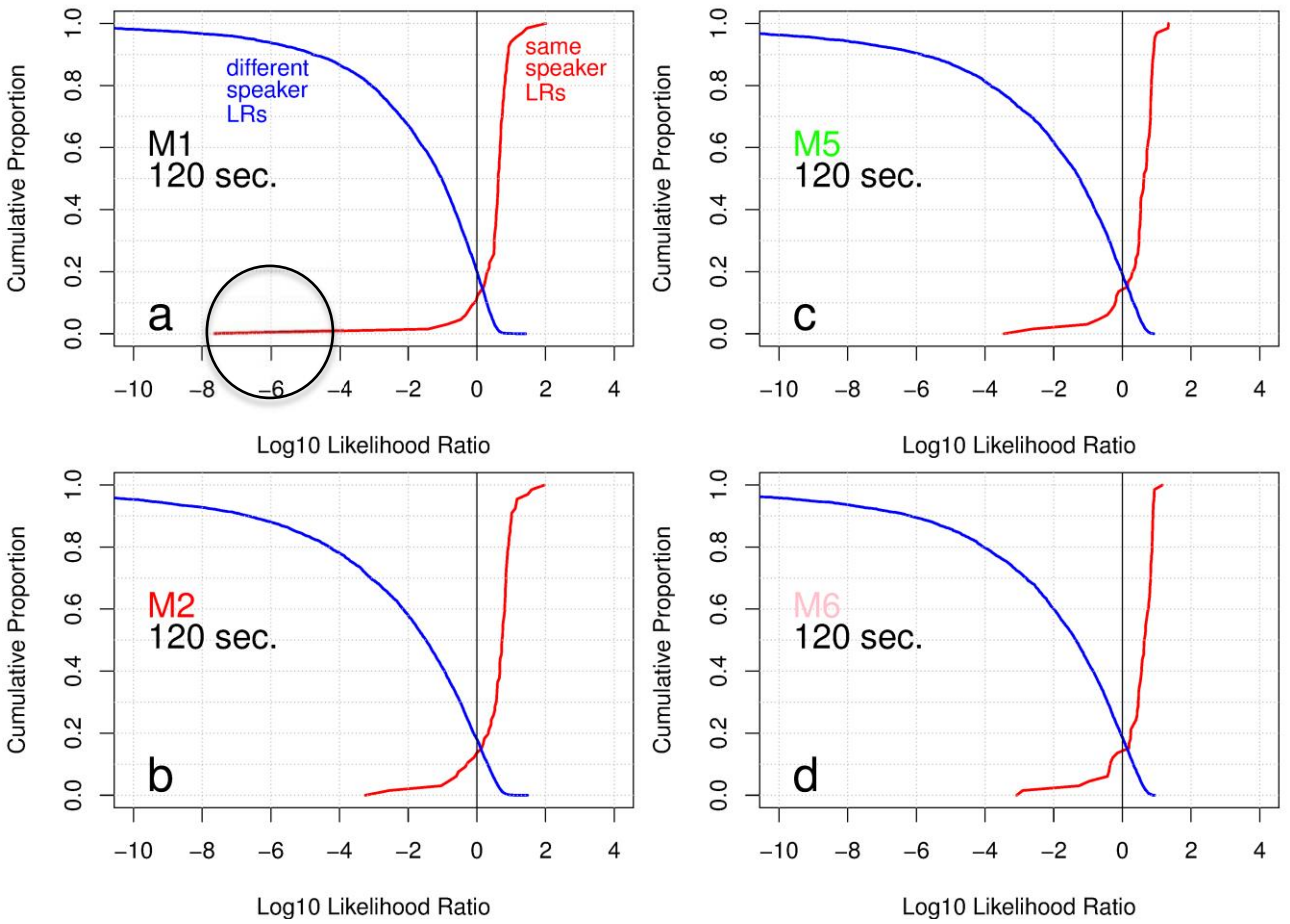


Figure 3: Tippett plots of M1, M2, M5 and M6 for the sample size of 120 sec. Red and blue curves = same speaker and different speaker LRs, respectively. The circle indicates large counter-factual same speaker LRs.

## 7. Conclusions

As far as the results of the current study are concerned, the positive effect of the multiple sets of feature vectors was not consistently observed, nevertheless, it was also pointed out i) that the multiple sets of feature vectors appear to perform better than the benchmark (M1) when the sample size is large (e.g. 100 sec. or longer) and ii) that M2 (two sets of feature vector) consistently performed better than M1, except when the sample size is 80 sec. An interpretation of the result of the present study is that the caseworker should try some different sampling procedures to see how the derived LR fluctuates.

It is theoretically interesting to know that the discriminability was relatively comparable across the different sample sizes (M1-M6), whereas the calibration loss largely fluctuated depending on the sample size if the analysis windows are not overlapped (M1-M4). The effect of the overlapped analysis windows is evident in that the calibration loss becomes fairly stable.

It is also interesting to see how the performance of the FVC system will improve when the F0-based LRs are fused with the LRs obtained from more powerful filter-based features (e.g. formant frequencies).

## 8. Acknowledgements

I would like to thank the two anonymous reviewers for their valuable comments.

## 9. References

[1] Y. Kinoshita, S. Ishihara, and P. Rose, "Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition," *International Journal of Speech Language and the Law,* vol. 16, pp. 91-111, 2009.

[2] S. Ishihara, "The Effect of the Within-speaker Sample Size on the Performance of Likelihood Ratio Based Forensic Voice Comparison: Monte Carlo Simulations," in *Australasian Language Technology Association Workshop 2013*, Brisbane, Australia, 2013, pp. 25-33.

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

[4] B. Robertson and G. A. Vignaux, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester: Wiley, 1995.

[5] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *the 2nd International Conference of Language Resources and Evaluation*, Athens, Greece, 2000, pp. 947-952.

[6] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language,* vol. 20, pp. 230-275, Apr-Jul 2006.

[7] B. Nair, E. Alzqhoul, and B. J. Guillemin, "Determination of likelihood ratios for forensic voice comparison using Principal Component Analysis," *International Journal of Speech Language and the Law,* vol. 21, pp. 83-112, 2014.