LR estimation using long term F0 as a parameter: good, bad or useless? Initial investigation using Japanese data

Yuko Kinoshita

School of Language and International Education,
University of Canberra
Yuko.Kinoshita@canberra.edu.au

Abstract

This paper investigates the validity of LR estimation for long-term F0 using Aitkin (1995)'s formula. Although this formula has been developed to estimate the LR of reflective index of glass fragments, previous studies such as Kinoshita (2001) and Rose, Osanai, and Kinoshita (2003) have shown that Aitkin's formula can be applied to speech data. The experiments in this study revealed, however, that it is not suitable for LR estimation when the long-term F0 is used as the speaker identification parameter. Other avenues for using F0 are suggested.

1. Introduction

The "Likelihood Ratio" (LR hereafter) has become an indispensable concept in forensic science in recent years, especially in the evaluation of evidence to be presented to legal proceedings. In Kinoshita (2001), the author has performed an LR-based discrimination for forensic speaker identification, using Aitkin's LR estimation formula with formants as the parameters. The discrimination test produced a success rate of about 96%, showing Aitkin's formula works well with formant data even though it was not developed for evaluation of speech data. In this study, the same method is applied to the long-term F0, which has also been noted as a potential speaker identification parameter (for instance, see Sambur 1975, Nolan, 1983:124, Jiang 1996). This paper focuses on investigating the usability of the application of Aitkin's formula to the long-term F0, as a preparatory process for a larger study.

F0 can vary dramatically from occasion to occasion within a speaker, and many factors are known to affect F0. Emotional changes, speaking style, the noisiness of the environment, and whether or not the person is on the phone are some of the factors reported to cause F0 variations (Maekawa 1998, Watanabe 1998, Boss 1996, Elliott 2000, French 1994, Hirson, French, & Howard, 1995). F0 is thus not always an acceptable parameter to use — it is necessary to first establish the comparability between samples through auditory analyses by a forensic speaker identification expert. F0 is, however, relatively robust against poor recording quality and is also an easier parameter to measure than others, such as formants. Because of its large within-speaker variability, F0 is not expected to be a very powerful parameter by itself, but it may still be useful enough as an additional piece of information on the speaker's identity. It seems thus unjust to neglect even investigating its potential.

This study investigated the potential of F0 as a forensic speaker identification parameter — and in particular whether or not Aitkin's formula produces useful LR estimates — by conducting two experiments.

The first experiment investigated the potential range of LR estimates calculated using the formula from Aitken, (1995:181), to analyse speech data using long-term F0. This experiment artificially manipulated the hypothetical F0 means, and the standard deviations (SD hereafter), within bounds representing typical Japanese speech data.

In the second experiment, tests were performed using actual speakers in order to find out how well Aitkin's formula works with this parameter. The long-term F0 of spontaneous speech spoken by 12 male Japanese speakers was compared using the same estimation formula as the first experiment.

2. Likelihood ratio

2.1. Expression of LR

The LR is the probability that the evidence would occur if our assertion is true, divided by the probability that the evidence would occur if the assertion is not true (Robertson & Vignaux, 1995:17). This can be expressed as follows:

$$LR = \frac{\mathrm{P(E\,|\,H)}}{\mathrm{P(E\,|\,\overline{H})}} \quad (1)$$

The LR is always expressed as a positive ratio. The LR will be larger than 1 when the given evidence supports the hypothesis, whereas it will be smaller than 1 when the evidence doesn't support the hypothesis. The relative distance of the LR from 1 measures the strength of the evidence.

It is important to note here that although the direction that LR points to changes at 1, LR is not a binary expression of truth. In other words, it does not answer the question "Are these two samples from the same speaker?" but expresses the strength of evidence in a continuous scale.

We still do not have a clear answer to an important question: how should one interpret an LR in evidence, for example the assertion "these two recordings are five times

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

more likely to come from the same person than not”? Champod and Evett (2000) propose, however, a verbal scale of strength of evidence that corresponds to LRs as below.

0.1–10		limited strength
10–100	0.1–0.01	moderate
100–1000	0.001–0.0001	moderately strong
1000–10000	0.001–0.0001	strong
>10000	<0.0001	very strong

2.2. Likelihood ratio estimation with Aitkin’s formula

Unlike with categorical data, estimating the probability of the evidence supporting — or not supporting — a hypothesis with continuous data requires a mathematically complex calculation. The formula used in this study is from Aitken (1995:181). This formula was originally developed for the evaluation of reflective indices of glass fragments, not for speech. The nature of variation in speech data is more complex than that of glass fragments, as speech is known to differ from one occasion to another, even when uttered by a single speaker. The glass fragments, on the other hand, do not change their quality over time. Since Aitkin’s formula was for glass fragments, it did not have to take into account this occasion-dependent within-speaker variation. This can be a problem; however, previous studies based on this formula indicate that it works sufficiently well with parameters such as formants and cepstrum (for instance, see Kinoshita 2001, Rose, Osanai, & Kinoshita 2002). Because of this, and also currently having no alternative available, it seems worthwhile to test the applicability of Aitkin’s formula in the current context.

$$V \cong \frac{\tau}{a\sigma} \times \exp\left\{-\frac{(\bar{x}-\bar{y})^2}{2a^2\sigma^2}\right\} \times \exp\left\{-\frac{(w-\mu)^2}{2\tau^2} + \frac{(z-\mu)^2}{\tau^2}\right\} \quad (2)$$

In this formula, the estimation of LR takes seven values into account: 1) Number of measurements of criminal samples (m), 2) Mean of criminal samples ($\bar{x}$), 3) Number of measurements of suspect samples (n), 4) Mean of suspect samples ($\bar{y}$), 5) Variance of criminal and suspects samples (σ^2), 6) Overall mean of population (μ), 7) Overall variance of population (τ^2). Furthermore, *z, w,* and a^2 are derived using the values above as:

$$z = (\bar{x} + \bar{y}) / 2, w = (m\bar{x} + n\bar{y}) / (m + n), \quad (3)$$
$$a^2 = 1 / m + 1 / n.$$

It is important to distinguish the LR from the estimate of the LR. The value produced using formula above is not necessarily an LR per se — it is an estimate of the LR. This paper investigates how well this formula works in the production of an LR estimate using long-term F0.

2.3. Likelihood ratios in forensic science

In section 2.1and 2.2, the concept of LR was described. Why is it so important in the current context? This section discusses how LR is relevant to forensic science.

Robertson & Vignaux (1995:21) give two reasons why LRs should be used for evidence evaluation and presentation. Firstly, the majority of evidence submitted to the court is by nature only indicative, not determinative. It is thus not likely that experts will be able to deliver conclusive remarks. The other reason is a result of the expert’s role in the legal system: they are not in a position 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 task of experts is to evaluate the likelihood of observing given evidence when a particular hypothesis — usually the prosecution’s — is correct versus when it is incorrect: ie, an LR.

In addition to appropriateness to the legal system, LRs have another feature in evidence presentation: a single LR can be produced for several pieces of evidence. It is very easy to combine multiple LRs by applying Bayes’ theorem, and thus the overall estimation is produced very straightforwardly. This is a significant feature, as most court cases involve multiple pieces of evidence. This becomes even more significant in the evaluation of speaker’s 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 a sufficient number of parameters in order to evaluate speech evidence (for instance, see Kinoshita, 2002), and the use of LRs and Bayes’ theorem facilitates this process by providing us with a simple and systematic approach.

3. Data

3.1. Population mean and SD

In Aitkin’s LR estimation formula, population mean and SD play significant roles. This is because the evaluation of the samples is based not only on the similarity of two samples, but also on their typicality against the population. Two samples being very similar does not mean much if those two are typical values in the population: for example, if there are many speakers who can conceivably produce those samples. If those samples are located on the outskirts of the population distribution, on the other hand, their similarity has more significance. The experiments in this paper use data on population mean and SD from Furui’s unpublished data (Furui, personal communication). Furui’s laboratory at the Tokyo Institute of Technology sampled long-term F0 from spontaneous speech produced by 90 male Japanese native speakers included in “The Corpus for Spontaneous Speech” database. As a result, 135Hz and 26.5 are obtained as the population mean and SD, respectively. Hollien and Jackson report a mean F0 of 123.3 Hz in sampling from the spontaneous speech of 157 male English speakers’ (quoted in Backen, 1996:155). The mean F0 of 135 Hz for Japanese speakers is notably higher than Hollien and Jackson’s data. There are many variables that could be the causes of this difference, but part of it may be attributed to differences in physique, such as the size of vocal cords, between typical speakers of English and Japanese.

3.2. Data for Experiment 2

3.2.1. Informants

For the second experiment, 12 male native speakers of Japanese were recorded. Their age ranged 20–36 at the time of recording. In the recording sessions, the informants performed three tasks, which were designed to elicit spontaneous speech; the data obtained from the second task were used for this paper. In this task, an information sheet on four people was given to the informants. This described four people's jobs, personalities, and favourite foods. The informants were then requested to explain what kind of person each of those four people is, referring to the given information.

The recording was carried out in the studio of the Phonetics laboratory at the Australian National University, and two recording sessions were held for each speaker, separated by two weeks. The sequence of tasks was performed twice in each session. As a result, each speaker produced eight utterances for each recording session (describing 4 people * 2 repeats), 16 utterances in total. The duration of recordings varied speaker to speaker from 32 seconds to 74 seconds per recording session, as each utterance was completely spontaneous.

3.2.2. Measurements

The recordings were digitised at 16 kHz and analysed with Praat. The analysis range was set at 40-300Hz, and this set the sampling frame at 18.75 milliseconds.

The results of these measurements are summarised in Table 1 and 2.

	Session 1		**Session 2**	
speaker	**mean**	**SD**	**mean**	**SD**
aa	125.0	22.7	119.6	25.2
ha	111.2	24.0	119.0	25.0
jn	128.3	16.9	124.9	15.3
ka	117.1	22.6	111.1	27.6
kf	124.8	30.5	110.3	25.8
mm	125.1	23.3	119.5	24.7
mn	117.2	18.3	129.9	22.1
mo	105.3	21.1	105.7	22.8
tn	97.7	22.0	99.4	16.2
ts	112.2	19.3	108.2	17.3
ty	107.4	25.4	113.7	19.4
yh	114.2	11.1	111.8	13.2

Table 1: Mean and SDs for each speaker.

Further, the means and SDs for all speakers are summarised in table 2 below.

Session 1		**session 2**		**all**	
Mean	**SD**	**mean**	**SD**	**mean**	**SD**
115.5	21.4	114.4	21.2	114.9	21.3

Table 2: Mean and SDs for all speakers.

Compared to the population mean of 135Hz quoted from Furui, the mean F0 of 114Hz here is considerably lower. The reason for this difference is far from clear; however, differences in speech style may be a partial explanation. The informants for this study were recorded through interviews with the author, while samples from the speech database analysed by Furui were recorded during paper presentations at conferences. Talking to a large audience and being in a one on one interview would produce different styles of speech, and F0 is known to rise as a speaker tries to speak more loudly due to the increase of the subglottal pressure (Lehiste 1970:56).

4. Testing and results

4.1. Experiment 1

4.1.1. Manipulation of means

In this experiment, LRs were estimated with Aitkin's formula (described in section 2.2), simulating real-world data by setting a plausible range of means and SDs, in order to find out what range of estimated LRs were calculated with long tem F0.

Two data sets were simulated: a hypothetical criminal set, and a hypothetical suspect set. First of all, LR was calculated with the means of hypothetical criminal samples being set at 90 Hz, 100 Hz, 110Hz, 120Hz, 130Hz, 140Hz, 150 Hz, and 160 Hz. The mean for the hypothetical suspect's mean was altered from 85 to 160 Hz, in 1Hz increments. This F0 range was determined referring to the data gathered for the second experiment in this study and Sambur (1975:181), in which he reports that speakers can be categorised in low (around 100 Hz), mid (125 Hz) and high (160 Hz) groups based on their long-term F0. The SD was kept at 21 for both criminal and suspect's samples at first. As the mean SD of the 12 speakers sampled for this study was 21.3, SD 21 seemed like an appropriate starting point for the experiment. The population mean and SD were quoted from Furui, thus they were set at 135.7Hz and 26.4.

Table 3 below summarises the relevant parts of the LR calculation results. As any suspect's mean which were more than 5 Hz distant from the criminal sample mean produced extremely small LRs, Table 3 presents the LRs produced with the suspect's mean sample which was within ±5Hz range of criminal sample mean. "CRIM M=" shows where the criminal sample mean was set, and the columns "S. M" indicate the mean of the suspect samples. Thus, this table shows, for instance, that when the criminal sample mean was 90 Hz and the suspect sample mean was 86 Hz, the LR for that comparison was 2.2x10^{-04}. The shaded columns indicate those combinations which produced LRs greater than 1.

CRIM M=90		CRIM M=100		CRIM M=110		CRIM M=120	
S. M	LR	S. M	LR	S. M	LR	S. M	LR
85	1.1E-07	95	5.8E-08	105	9.5E-08	115	6.8E-08
86	2.2E-04	96	1.2E-04	106	1.4E-04	116	9.9E-05
87	8.1E-02	97	4.4E-02	107	3.9E-02	117	2.8E-02
88	5.5E+00	98	3.0E+00	108	2.2E+00	118	1.6E+00
89	6.8E+01	99	3.8E+01	109	2.4E+01	119	1.8E+01
90	1.5E+02	100	8.6E+01	110	5.4E+01	120	4.0E+01
91	6.4E+01	101	3.6E+01	111	2.4E+01	121	1.8E+01
92	4.8E+00	102	2.7E+00	112	2.0E+00	122	1.5E+00
93	6.7E-02	103	3.8E-02	113	3.5E-02	123	2.6E-02
94	1.7E-04	104	9.7E-05	114	1.2E-04	124	9.0E-05
95	7.8E-08	105	4.5E-08	115	7.9E-08	125	6.1E-08

CRIM M=130		CRIM M=140		CRIM M=150		CRIM M=160	
S. M	LR	S. M	LR	S. M	LR	S. M	LR
125	5.6E-08	135	5.4E-08	145	2.3E-08	155	2.9E-08
126	8.2E-05	136	7.9E-05	146	4.8E-05	156	6.1E-05
127	2.4E-02	137	2.3E-02	147	1.8E-02	157	2.4E-02
128	1.4E+00	138	1.3E+00	148	1.3E+00	158	1.7E+00
129	1.5E+01	139	1.5E+01	149	1.7E+01	159	2.2E+01
130	3.4E+01	140	3.4E+01	150	4.0E+01	160	5.3E+01
131	1.5E+01	141	1.5E+01	151	1.7E+01	161	2.3E+01
132	1.3E+00	142	1.3E+00	152	1.4E+00	162	1.8E+00
133	2.3E-02	143	2.3E-02	153	2.0E-02	163	2.6E-02
134	7.9E-05	144	8.1E-05	154	5.2E-05	164	7.0E-05
135	5.4E-08	145	5.5E-08	155	2.5E-08	165	3.4E-08

Table 3: LRs produced with the various criminal and suspect sample means. The SD for both samples was set at 21.

The highest LR was 1.5x10^2 (154.0) where the criminal and suspect's sample means were both 90Hz. According to aforementioned Champod and Evett (2000)'s scale, 1.5E+02 falls in the category of moderately strong evidence. No matter where the criminal sample mean was located in the population distribution, the LR was generally very small unless the criminal and suspect's sample means were within 1Hz.

The results revealed that the LR estimates for the long-term F0 tend to be extremely small. Regardless of the criminal sample mean, an LR above 1 was produced only when the criminal and suspect's sample means were within 2Hz. As described in section 2.1, Champod and Evett (2000)'s verbal scale rates LRs at 1000-10000 (or 0.001 - 0.0001) as strong evidence and anything beyond that as very strong evidence. Although the possible range for a reliable LR has not been established, comparison to the this verbal scale suggests that the LR estimates obtained in this experiment may be out of scale. For example, with the mean of the criminal sample at 90 Hz, and the suspect's sample mean set at 96Hz, the LR was 6.6 x 10^{-12}. With the suspect's sample mean larger than this, the LR was even smaller. This would be a concern in actual forensic speaker identification, as the data collected for this study (presented in Table 1) has shown that the mean F0 difference between two sets of recordings from single speaker can easily exceed 6Hz. Seven speakers out of 12 had 6Hz or more difference between two recording sessions in this study.

4.1.2. Manipulation of SD

Next, the effect of the SDs was looked into, by manipulating the SDs of both criminal and suspect's samples. The SD was moved among the values of 11, 15, 20, 25, and 30. This range was decided referring to the SDs obtained from the 12 speakers recorded for this study. Since the LR estimate becomes extremely small when the means of criminal and suspect samples are more than 5 Hz apart, the suspect sample means used for the calculation this time were limited to ±5 Hz range from the criminal sample means. Thus, in this section, the suspect's sample mean was altered only between 85 and 95 Hz when the criminal sample mean was set at 90Hz, and the SDs for criminal and suspect's samples were any combination of 11, 15, 20, 25, and 30. A part of the results is presented in Table 4 and Table **5**. These are the results obtained when the criminal mean was 90Hz, and suspect's mean was between 85 and 95Hz. Each row and column indicates a different SD used for LR estimation. The highest LR estimates for each SD combination were presented in Table 4. Table 5 presents how small the differences in means have to be in order to produce LR above 1 for each SD combination. The numbers "11, 15, 20, 25, 30" in the first row and column indicate the size of SD used for LR estimation.

	11	**15**	**20**	**25**	**30**
11	294.0				
15	245.9	215.6			
20	200.4	182.9	161.7		
25	167.4	156.9	142.8	129.3	
30	143.1	136.3	126.8	117.1	107.8

Table 4: The highest LR estimates obtained were presented (the criminal mean was 90Hz, and the suspect's mean was 85-90Hz).

	11	**15**	**20**	**25**	**30**
11	1Hz	1Hz	1Hz	2Hz	2Hz
15	1Hz	1Hz	2Hz	2Hz	2Hz
20	1Hz	2Hz	2Hz	2Hz	2Hz
25	2Hz	2Hz	2Hz	2Hz	3Hz
30	2Hz	2Hz	2Hz	3Hz	3Hz

Table 5: The range of the difference (±) between the criminal and suspect's means which produced LR estimates greater than 1 (the criminal mean was 90Hz, and suspects mean was 85-90Hz).

When the means are very close to each other, naturally the smaller SDs produce the higher LR estimates. For instance, Table 5 shows that when the both criminal and suspect's samples had the mean of 90Hz and the SD of 11, the LR was 294.0, although this is a highly unlikely situation in reality.

The tables above have shown that Aitkin's formula can produce high LR estimates, but the criminal and suspect's means have to be very close to each other, within ±3Hz at most. Given the large variability of F0, this result seems to suggest that Aitkin's formula does not work very well with long-term F0, which contains potentially a large occasion-to-occasion variation.

It also appears that the distribution of LR estimates was biased towards values below 1. This may be attributed to the fact that Aitkin's formula does not take occasion-to-occasion-variation into account. This formula thus does not recognize the possibility of one speaker varying from one occasion to the other. The variations between samples are perhaps regarded as between-speaker variation more often than they actually are, consequently producing more LRs smaller than 1, which point to the recordings coming from two separate individuals.

4.2. Experiment 2

4.2.1. Result of likelihood estimation

The experiment with the hypothetical F0 distribution revealed that long-term F0 tended to produce small LR estimates, and perhaps that the entire distribution of LR in this parameter are pushed towards values below 1 because of the nature of the formula. In this section, a test of the plausibility of the LR estimation is performed to see how the characteristics of long-term F0 discussed above are realised in an actual speech data scenario. Here, actual F0 data elicited from 12 speakers' natural speech were used as testing samples. Table 6 summarises LR estimates produced from same-speaker comparisons. The comparisons were made between two recording sessions for each speaker. As detailed earlier, the LR is a ratio: the probability of observing the evidence if the speech came from the same person divided by the probability of observing the evidence if the speech did not come from the same person. The LR estimates for the same speaker comparison are therefore expected to be generally greater than 1. The shaded cells indicate where the LR was above than 1.

speaker	LR
aa	6.3E-08
ha	2.4E-22
jn	3.4E-05
ka	7.0E-06
kf	2.8E-32
mm	3.5E-10
mn	1.2E-44
mo	5.2E+01
tn	1.3E+00
ts	1.3E-06
ty	3.1E-10
yh	1.7E-04

Table 6: LR estimates for the same speaker comparisons

The results for this testing were astonishingly poor. Only two speaker samples, speakers *mo* and *tn*, produced any estimated likelihood of being the same speaker pair. The rest produced very low LRs. While the number of pairings is very small, the average estimated LR for this group of pairs was not a plausible LR.

Next, some of the LR estimates for different speaker pairs are presented. Since each speaker had two recording sessions, one pair of speakers produced four different comparisons (e.g. speaker *aa* session 1 / speaker *ha* session 1, speaker *aa* session 1 / speaker *ha* session 2, speaker *aa* session 2 / speaker *ha* session 1, speaker *aa* session 2 / speaker *ha* session 2).

The LR estimates for the different-speaker comparisons are expected to be generally less than 1. The shaded cells indicate where the LR estimate was larger than 1.

The comparisons of different speakers also produced implausible LR estimates. 264 LR estimates were produced for different speaker comparisons. 31 of them were greater than 1.

speaker	aa	ha	Jn	ka	kf	mm
ha	2.7E-65					
ha	2.3E-13					
ha	1.7E-19					
ha	2.9E+01					
jn	1.0E-03	5.8E-114				
jn	3.6E+01	2.3E-60				
jn	2.5E-25	9.7E-35				
jn	5.0E-15	1.5E-10				
ka	2.7E-18	5.8E-09	1.5E-46			
ka	1.5E-41	3.9E+01	1.5E-77			
ka	8.4E-01	2.9E+00	5.9E-34			
ka	4.5E-11	1.4E-11	6.9E-37			
kf	3.0E+01	2.8E-45	1.6E-02	1.8E-11		
kf	1.5E-49	2.9E+01	2.0E-92	2.9E-08		
kf	5.4E-05	2.2E-09	4.1E+01	2.4E-27		
kf	2.3E-15	2.4E-16	7.4E-48	2.9E+01		
mm	3.7E+01	2.9E-68	2.6E-03	4.3E-19	2.9E+01	
mm	2.3E-09	1.6E-21	5.9E-28	8.8E-01	5.9E-06	
mm	1.5E-08	1.2E-14	4.5E+01	7.3E-43	7.7E-65	
mm	3.6E+01	3.4E+01	5.3E-10	3.1E-14	1.6E-22	
mn	1.6E-17	3.6E-09	1.2E-50	3.8E+01	4.6E-10	3.9E-11
mn	1.2E-06	3.2E-100	3.3E+00	4.2E-43	1.5E-04	4.0E-108
mn	1.3E+00	4.7E+00	3.0E-34	2.2E-06	2.3E-10	7.0E-28
mn	1.4E-30	9.1E-44	9.0E-09	2.0E-82	6.4E-112	3.2E+00

Table 7: LR estimates for the different speaker comparisons

The smallest LR estimates were over a hundred orders of magnitude smaller than the smallest LR estimates of formants presented in Kinoshita (2001). Kinoshita calculated the LR estimates using six different formant / vowel combinations. The formants were sampled from natural speech spoken by ten male Japanese speakers, and the LR estimates were produced for 180 different speaker comparisons and 90 same speaker comparisons. In that experiment, for the same speaker comparison, the largest LR estimate obtained was 110.7, which is a moderately strong evidence against hypothesis according to the Champod and Evett (2000)'s verbal scale, and the smallest LR was 0.01, which is classified as the strong evidence. For the different-speaker comparison it was 104.4 and $1.2\text{x}10^{-114}$. This $1.2\text{x}10^{-114}$ seems absurdly small, given that anything smaller than 0.0001 is supposed to indicate a very strongly opposition to the hypothesis. However the smallest value in this study was far smaller than this.

It is not possible to make a straightforward comparison, but it seems that the LR estimates for this study are considerably smaller. This may mean that formants are less susceptible to occasion-to-occasion variation and thus less influenced by the shortcomings of the formula used in this study.

The second experiment made it very clear that Aitkin's formula is not suitable for estimating LRs for long-term F0. As mentioned earlier, LR is not a binary expression of truth but presents the strength of the evidence in continuous scale. In other words, LR 5 does not mean that the suspect and criminal are the same speaker, but rather that it is five times more likely to be so than not. Therefore, there is always a possibility that the LR in question is pointing the "wrong" direction. Since speech is the product of an extremely complex process and has a very high variability within a speaker, it would not be practical to ask for an LR estimation formula which works perfectly. However, what we need is is to have a formula which does not point to the wrong direction with strong confidence. In this study, the comparisons of the same speakers have produced LR estimates > 1 for two out of 12 comparisons. This is alarming, but the more significant problem lies in the fact that the ten same-speaker comparisons all produced extremely low LRs, much <0.00001 which is the level of "very strongly against." Such extreme LR estimates can very easily overpower the whole evaluation of speech evidence and make it point to the wrong direction, even when various other pieces of evidence were suggesting a different conclusion. This is clearly the situation that we must prevent.

5. Conclusion

This paper investigated the potential application of Aitkin's LR estimation formula to the long-term F0.

First of all, the potential range of LR estimates produced with Aitkin's formula was investigated based on hypothetical F0 distributions. It was found that this analysis produces exceedingly small LR estimates. This raised doubts as to the appropriateness of applying Aitkin's formula to long-term F0. This doubt was confirmed by the second experiment of this research, in which LR estimates for actual speech data comparisons were produced. Many comparisons between the same speakers produced extremely small LR estimates. This, together with some absurdly small LR estimates, clearly indicates that a better model to estimate the LR for F0, and quite possibly speech data in general, is necessary.

As a future task, while searching for a better LR estimation formula, there is an alternative approach for the use of long-term F0. Sambur (1975:181) reports that speakers can largely be categorised into three groups — low, mid, and high — based on their long-term F0; and that speakers rarely cross the category boundaries although they can vary radically within the range of their category. It may thus be possible to produce an LR based on the speakers' category. Given that this is a very broad classification, it cannot be expected to produce high LR estimates, but it may produce useable estimates of high reliability. This may be a worthwhile investigation to pursue.

6. Acknowledgement

I would like to thank Professor Furui at the Tokyo Institute of Technology, who responded to my inquiries on population mean and standard deviation by kindly providing me with the data used in this study.

7. References

Aitken, C. G. G. (1995). *Statistics and the Evaluation of Evidence for Forensic Scientists*: Wilely.

Backen, R. J. (1996). *Clinical Measurement of Speech and Voice*. San Diego: Singular Publishing Group, Inc.

Boss, D. (1996). The problem of F0 and real-life speaker identification: case study. *Forensic Linguistics, 3*(1), 155-159.

Elliott, J. (2000). *Comparing the acoustic properties of normal and shouted speech: a study in forensic phonetics*. Paper presented at the The Eighth Australian International Conference on Speech Science ad Technology, Canberra.

French, P. (1994). An overview of forensic phonetics with particualr reference to speaker identifidation. *Forensic Linguistics, 1*, 169-181.

Hirson, A., French, P., & Howard, D. (1995). Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics. In *Studies in General and English Phonetics*: Routledge.

Jiang, M. (1996). Fundamental frequency vector for a speaker identification system. *Forensic Linguistics, 3*(1), 95-106.

Kinoshita, Y. (2001). *Testing realistic forensic speaker identification in Japanese: A likelihood ratio based approach using formants.* Unpublished PhD, The Australian National University, Canberra.

Kinoshita, Y. (2002). *How small can it get? Forensic speaker identification as a function of parameter number.* Paper presented at the The ninth Australian International Conference on Speech Science and Technology, Melbourne.

Maekawa, K. (1998, 30th November 4th December). *Phonetic and phonological characteristics of paralinguistic information in spoken Japanese.* Paper presented at the The 5th International Conference on Spoken Language Processing, Sydney.

Robertson, B., & Vignaux, G. A. (1995). *Interepreting Evidence*. Chichester: Wiley.

Rose, P. J., Osanai, T., & Kinoshita, Y. (2002). *Strength of forensic speaker identification evidence: multispeaker formant and cepstrum-based segmental discrimination with a bayesian likelihood* ratio *as threshold.* Paper presented at the The ninth Australian International Conference on Speech Science and Technology, Melbourne.

Sambur, M. R. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-23*(2), 178-182.

Watanabe, T. (1998). Japanese pitch and mood. *Nihongakuho, Osaka University, 17*, 97-110.