
Supawan Pingjai
Chunichi Ishihara

Abstract
This paper discusses the very first forensic voice comparison (FVC) using the Likelihood Ratio-based approach in the Thai language. FVC concerns the task of comparing two or more speech samples in order to assist the trier of the fact in determining whether two or more speech samples (usually of the suspect(s) and the offender) came from the same speaker or from different speakers. This experiment employs speech acoustics (F0 and duration) from 10 male speakers whose speech samples were elicited in a citation manner. In one test configuration 96% of same-speaker speech data were correctly identified as coming from the same speaker. This suggests that FVC works reasonably well in the Thai language.

Key words: Forensic Voice Comparison, Likelihood Ratio, polynomial, Equal Error Rate, Tippett plot.

บทคัดย่อ
บทความนี้นำเสนอผลการศึกษาการเปรียบเทียบเสียง (Forensic Voice Comparison) โดยใช้ลักษณะเชิงกลศาสตร์ในภาษาไทย ด้วยหลักการ Likelihood Ratio การเปรียบเทียบเสียงนี้ช่วยสนับสนุนกระบวนการยุติธรรมในการคำนวณความเป็นไปได้ว่าเสียงสองเสียงที่มาจากกว่า (มักจะเป็นเสียงของผู้ต้องสงสัยและผู้กระทำผิด)ว่าเป็นมาจากบุคคลเดียวกันหรือต่างบุคคลกัน ซึ่งการศึกษาการเปรียบเสียงในภาษาไทยนี้ใช้หลักการคือ ความถี่มูลฐาน (F0) และระยะเวลาในการเปล่งเสียง (duration) ที่เปล่งโดยผู้บอกภาษาพื้นชาติจำนวน 10 คน จากผลการทดลองแสดงให้เห็นว่าการคำนวณความเป็นไปได้ว่าเสียงสองเสียงมาจากบุคคลเดียวกันให้ความความถูกต้องถึง 96%

คำสำคัญ: การเปรียบเทียบเสียง, Likelihood Ratio, สัมประสิทธิ์, ความผิดพลาดที่เท่ากัน, กราฟความเชื่อมั่น
1. Introduction

With the advancement of technology in telecommunication, voice communication via telephone and the internet are common means for people to communicate amongst themselves. It is also a conceivably common means for perpetrators to communicate about matters related to crimes such as robbery, kidnapping, cheating, fraud, etc.

Assuming a scenario where there is an incriminating bomb threat, how well can forensic phoneticians determine the likelihood that this voice belongs to a particular suspect based on a prior voice recording? This paper introduces the Likelihood Ratio approach for evaluating the strength of evidence under two competing hypotheses (1 and 2); namely, (1) two voice samples come from the same speaker, or (2) two voice samples come from different speakers. This paper presents the first study of this kind in the Thai language.

1.1 Expressing the outcome: strength of the evidence

In order to tender evidence to the court, the piece of evidence is primarily required to be “relevant” to the case (Robertson and Vignaux 1995: 11). In practice, forensic scientists seldom find evidence which always occurs when the fact they are trying to prove exists (i.e. true) and, on the other hand, never occurs when the fact they are trying to prove does not exist (i.e. not true). This type of evidence is called the “ideal evidence”. It is typical, however, to find evidence which is not ideal but can be shown to be more likely to occur when the fact to be proven is true than when it is not. We call this type of evidence a “good” or “strong” evidence (Robertson and Vignaux 1995: 12-13). The purpose of this section is to generate a means of assessing the strength of the evidence gathered. Firstly, with forensic speech experts aiding the court or the police in coming to a conclusion about the origin of a set of speech samples, the expression the investigators may use to express the outcome could be written as follows:

$$p(H | E)$$  

**Formula 1**

where $p= \text{probability}$, $H = \text{hypothesis}$ and $E = \text{evidence}$, that is the probability of a certain hypothesis (H) given (indicated by “|”) the evidence (E).

This is called “the probability of a hypothesis given the evidence”, Rose (2002: 56). If Formula 1 is applied to a Forensic Voice Comparison (hereafter referred to as FVC) case, it is identical to saying “what is the probability of a hypothesis, such as “the speech samples are from the same speaker” (same speaker hypothesis $= H_{SS}$) or “the speech samples are from different speakers” (different speaker hypothesis $= H_{DS}$), given the evidence which is the amount of similarity/difference between the speech samples” (E), Rose (2002: 56).

However, such a formula, which identifies the probability of a hypothesis
given the evidence is problematic in that it appears to be legally and logically incorrect. By saying, for example, that the speech samples are more likely to come from the same speaker (Hss) given much similarity of the speech samples (E) would be to violate the duty of judges and juries by the forensic speech experts assuming their respective duties (Aitken 1995: 4; Evett 1991: 13-14; Rose 2002: 56-57; 2006b: 161-163). Furthermore, it is impossible to directly calculate the probability of a hypothesis given the evidence because the FVC experts cannot access and evaluate all of the evidence that is relevant to the case (Rose 2002: 56). The expert should instead report the probability of the evidence (as opposed to the hypothesis) given a hypothesis. In other words, when the opinion of the forensic-speech investigator is sought, the probability of the evidence (as opposed to the probability of the hypothesis), given two hypotheses (i.e. the hypothesis that two speech samples were produced by the same speaker or the same speaker hypothesis (Hss) vs. the hypothesis that two speech samples were produced by different speakers or the different speaker hypothesis (Hds)) should be quantified, and the ratio of the probabilities calculated in order to express the outcome of a FVC. This ratio is formulated in Formula 2, this is an expression for the strength of evidence or otherwise called the Likelihood Ratio (LR), Rose (2002: 58).

\[
\text{LR} = \frac{p(E|H_{ss})}{p(E|H_{ds})} \quad \text{Formula 2}
\]

where \( p \) = probability, \( E \) = evidence, \( H_{ss} \) = the same speaker hypothesis or the prosecution hypothesis, \( H_{ds} \) = the different speaker hypothesis or the defense hypothesis.

If this observed degree of similarity had it come from the same speaker is 90% in probability, thus \( p(E|H_{ss}) = 90\% \). And if this observed degree of similarity had it come from different speakers is 10% in probability, thus \( p(E|H_{ds}) = 10\% \). Then the strength of evidence or LR can be calculated as:

\[
\text{LR} = \frac{0.90}{0.10} = 9 \quad \text{Formula 3}
\]

This is equivalent to saying that this degree of similarity/difference between speech samples is nine times more likely observed if it had come from the same speaker than if it had come from different speakers.

As was explained above, the LR is the ratio of the probability of the evidence supposing the hypothesis is true (numerator) divided by that of the evidence supposing the hypothesis is false (denominator), Robertson and Vignaux (1995: 17). A single LR figure will result and this figure will tell us the strength of the evidence in support of the hypothesis (ibid.). The strength of evidence can be
measured by how much the LR is greater than or less than 1 (ibid.). That is to say, if the LR value is more than 1 then the evidence supports the hypothesis, if it is far greater than 1 then the evidence strongly supports the hypothesis. On other hand, if the LR is less than 1 then the evidence goes against the hypothesis. And if the LR is exactly 1, the strength of the evidence is neutral and the evidence is therefore useless for supporting either hypothesis. It is a common practice to use $\log_{10}LR$ in which the unity is 0, (Rose 2002; 2006b).

There are many formulae for estimating the LR (Lindley 1977; Reynolds et al 2000; Aitken and Lucy 2004). In this study, Aitken and Lucy’s (2004) formula for a multivariate likelihood ratio (MVLR) is used. The MVLR is explained in detail in the following section.

1.2 Combination of LRs
LRs are perfectly suited to testing how well same-speaker speech samples can be discriminated from different-speaker speech samples. The reasons why LRs are suitable in forensic science are that the evidence is naturally indicative (not determinative), so the forensic scientists cannot directly indicate the guilt of the defendant given the evidence (which is prohibited in most legal systems). Fact finding is the job of the judge or jury who has access to the rest of other evidence of the case in order to consider the probability of guilt (Rose 2006b: 162; Morrison: 2009: 300-301). This is aided by the fact that LR can be combined with LRs from other evidence via Bayes’ theorem. If the different pieces of evidence are independent, it is easy to derive the overall LR by multiplying the individual LR from each item of evidence together (Rose 2002: 60-61; Robertson and Vignaux 1995: 69-71). For example, the overall LR of 14 is derived by the LR of the first piece of evidence, say blood type, (7) multiplied by the LR of the second piece of evidence, say voice, (2), that is (7x2) =14.

LR framework was also shown to be an appropriate framework in FVC by many studies. Kinoshita (2001)’s empirical study shows that 90% of same-speaker speech samples are resolved correctly with LRs greater than 1 while 97% of different-speaker speech samples are resolved correctly with LRs less than 1. Kinoshita’s results conform to others’ studies such as Meuwly and Drygajlo (2001: 150) whose study shows that 86% of same-speaker speech samples have LRs greater than 1 and 86% of different-speaker speech samples have LRs less than 1. Similar results were also obtained in Spanish and American English studies of Gonzalez-Rodriguez et al. (2001) and Nakasone and Beck (2001) respectively.

1.3 Formula for calculating LRs
When the Likelihood Ratio (LR) is estimated based on speech samples, many of the acoustic parameters which describe the voices are correlated. Bayes’ theorem allows the LRs obtained from different sets of parameters to be combined,
however, they must not be correlated (Robertson and Vignaux, 1995: 70-71). The problem of estimating LRs from correlated variables was addressed by Aitken and Lucy when they formulated the multivariate LR formulae (MVLR) (Aitken and Lucy, 2004). Although the MVLR can only accommodate two levels of variance, it is a considerable improvement over univariate approaches. The MVLR formula can be written as follows:

**The MVLR formula**

\[
\text{numerator of MVLR} = (2\pi)^{\frac{p}{2}} |\mathbf{D}_1|^{\frac{1}{2}} |\mathbf{D}_2|^{\frac{1}{2}} |\mathbf{C}|^{\frac{1}{2}} \left( n_1^{-\frac{1}{2}} \mathbf{D}_1 + n_2^{-\frac{1}{2}} \mathbf{D}_2 + h^{-1} \mathbf{C} \right)^{\frac{1}{2}} \\
\times \exp \left\{ -\frac{1}{2} \mathbf{y}_l - \mathbf{x}_l \right\} \left[ n_1^{-\frac{1}{2}} (\mathbf{D}_1^{-1} + h \mathbf{C})^{-\frac{1}{2}} (\mathbf{D}_1^{-1} \mathbf{y}_l + h^{-1} \mathbf{C} \mathbf{x}_l) \right]
\]

\[
\text{denominator of MVLR} = (2\pi)^{\frac{p}{2}} |\mathbf{D}_1|^{\frac{1}{2}} |\mathbf{D}_2|^{\frac{1}{2}} |\mathbf{C}|^{\frac{1}{2}} \left( n_1^{-\frac{1}{2}} \mathbf{D}_1 + n_2^{-\frac{1}{2}} \mathbf{D}_2 + h^{-1} \mathbf{C} \right)^{\frac{1}{2}} \\
\times \sum_{i=1}^{n_1} \exp \left\{ -\frac{1}{2} (\mathbf{y}_i - \mathbf{x}_l) \right\} \left[ n_1^{-\frac{1}{2}} (\mathbf{D}_1^{-1} + h \mathbf{C})^{-\frac{1}{2}} (\mathbf{D}_1^{-1} \mathbf{y}_i + h^{-1} \mathbf{C} \mathbf{x}_i) \right]
\]

where \( \mathbf{D}_1, \mathbf{D}_2 = \text{within-speaker variance/covariance matrices}; n_1, n_2 = \text{number of replicates per speaker} \) 
\( m = \text{number of speakers in reference population}; p = \text{number of assumed correlated variables per speaker} \) 
\( h = \text{optimal smoothing parameter for kernel density} = \left( \frac{4}{(2p+1)} \right)^{1/2} m^{-1} (p+4) \) 
\( \mathbf{y}_l, \mathbf{y}_i = \text{offender, suspect variances} \) 
\( \mathbf{x}_l, \mathbf{x}_i = \text{within-speaker means of reference values} \) 
\( n_1^{-\frac{1}{2}} (\mathbf{D}_1^{-1} + h \mathbf{C})^{-\frac{1}{2}} (\mathbf{D}_1^{-1} \mathbf{y}_l + h^{-1} \mathbf{C} \mathbf{x}_l) \)

In the MVLR formula above, the numerator (1) estimates the similarity of the two samples under comparison while the denominator (2) estimates the typicality of the samples in a reference population (Alderman 2005: 25; Winter 2009: 24).

This formula is employed in the experiment as the parameters and variables used to describe speech that are being examined (the coefficients of a polynomial curve) are likely to be correlated to each other. The original MVLR formula assumes that the parameters of speech are normally distributed, however this assumption about speech cannot be made (Kinoshita 2001: 270; 295-297; Rose 2002: 321; Alderman 2005: 22). Instead of persisting with the normal distribution, the actual distribution (which may deviate from normality) can be modelled with a kernel density estimation (Aitken 1995). A kernel density estimation is a combination of normal distributions. Each normal distribution function is based on an observation of the model’s distribution. When they are combined, they give a more precise description of the actual distribution if it is not normal (Aitken 1995). Since deviation from normality can affect the accuracy of an LR estimate, the distribution of data in this experiment is modeled with a kernel density function. It is also important to mention here that only the background sample that can be modeled with a kernel density in this version of the MVLR. The reason is that the distributions of the suspect and offender samples may be too sparse to model with anything other than normal assumptions.
In order to evaluate the similarity of samples and their typicality to the rest of the population, the background or reference data must also be representative of the population (Alderman 2005: 25). In this experiment, we calculated the LRs in cross-validated manner, which means that a pair of speech samples under comparison was not included in the reference when the LR of the paired samples is estimated.

Finally, in order to aid the court of law to interpret LRs, verbal interpretations of the ranges of LRs have been proposed. The one that is used most frequently is the one used at the British Forensic Science Service on all biometric identification to indicate the amount of support for evidence (Gonzalez-Rodriguez et al. 2002: 174). The following verbal equivalents (Table 1) are proposed by Champod & Evett (in Rose 2002: 62). The scale of LRs, Log LRs and the strength of evidence they represent is presented in Table 1.

**Table 1: Verbal Equivalents of Likelihood Ratios (adapted from Rose 2002: 62).**

<table>
<thead>
<tr>
<th>Likelihood Ratio</th>
<th>Log10 Equivalent</th>
<th>Possible interpretation</th>
<th>Support for the prosecution hypothesis</th>
<th>Support for the defence hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 10 000</td>
<td>&gt; 4</td>
<td>Very strong...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000 to 10 000</td>
<td>3 to 4</td>
<td>Strong...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100 to 10 00</td>
<td>2 to 3</td>
<td>Moderately strong...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 to 100</td>
<td>1 to 2</td>
<td>Moderate...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 to 10</td>
<td>0 to 1</td>
<td>Limited...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 to 0.1</td>
<td>0 to -1</td>
<td>Limited...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1 to 0.01</td>
<td>-1 to -2</td>
<td>Moderate...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01 to 0.001</td>
<td>-2 to -3</td>
<td>Moderately strong...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.001 to 0.0001</td>
<td>-3 to -4</td>
<td>Strong...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 0.0001</td>
<td>&gt; -4</td>
<td>Very strong...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Table 1, the ranges of linear LR are presented in the first column, the corresponding Log10LRs in the middle column, and the possible verbal equivalents for both linear and Log LRs in the last column. Thus, when evidence with Log LR > 4 is tendered to the court, it suggests the “very strong” support for the hypothesis that the speech samples were more likely had they come from the same speaker (prosecution hypothesis) than had they come from different speakers. On the other hand, when the Log LR of -4 is presented to the court of law, this will count as “strong” support for the defence hypothesis that the speech samples are more likely to come from different speakers than from the same-speaker.

### 1.4 Equal Error Rate (EER)

The EER % measure indicates the level of misprediction in our model. This is a binary metric which indicates the percentage of incorrect predictions (LR < e), and the percentage of correct predictions (LR > e), where e is a LR threshold where the same-speaker and different-speaker LR probabilities are equal.
1.5 Tippet Plots

A Tippet plot (which is sometimes called reliability or probability distribution plot) is now the standard way of presenting results in FVC. Since it is difficult to explain a Tippet plot without looking at one, we will explain how to read it when presenting results using Tippet plots below. It is crucial to mention the importance for the forensic phoneticians in stating the probability of evidence (as opposed to the probability of the hypothesis) when the opinions of the forensic phoneticians are sought.

1.6 Choice of F0 as a parameter for this experiment.

Previous studies have successfully used the fundamental frequency (F0) (Kinoshita et al 2009) and formants of speech (Alderman 2005; Winter 2009) as discriminatory variables. These studies motivate the use of F0 in our research work.

2. Methodology

2.1 Corpus and elicitation

The twelve segmental combinations, of which syllable structure is a CVV, were created by permutating the three long vowels of /ii, aa, uu/ and the four consonantal phonemes of /p, pʰ, b, m/. These segmental combinations are given in Table 2. Five separate lists of these combinations were prepared for the five different tones. This is because with different tones, these sixty segmental combinations (12 segmental combinations x 5 tones) are lexically different words written differently in Thai orthography.

<table>
<thead>
<tr>
<th>pii</th>
<th>pʰii</th>
<th>bii</th>
<th>mii</th>
</tr>
</thead>
<tbody>
<tr>
<td>paa</td>
<td>pʰaa</td>
<td>baa</td>
<td>maa</td>
</tr>
<tr>
<td>puu</td>
<td>pʰuu</td>
<td>buu</td>
<td>muu</td>
</tr>
</tbody>
</table>

Table 2. The twelve segmental combinations

Ten native Thai speakers, all of whom were males, participated to this study as informants. Male speakers were chosen for this study because males feature more frequently in criminal cases (Statistical Yearbook Thailand 2007 (Special Edition)) that give rise to real cases of forensic voice comparisons. All speakers were students studying at the Australian National University at the time of recording, aged from 18 to 32. Another reason for choosing male informants for this study was that the male voice is generally less complicated than females according to Rose (personal communication).
2.2 Digitisation and measurement

The speech signals were directly recorded with *Audacity* using a professional microphone, and then stored as WAV files at a sampling rate of 16 kHz and 16 bit amplitude resolution. The audio speech waveform and the wide-band spectrogram of each target segmental combination was then extracted using *Praat*, and F0 plots were superimposed, using the standard F0 extraction algorithm (auto-correlation) with other standard settings. Figure 1 contains an audio speech waveform and a spectrogram of an example - [pʰaa mid-level tone] - uttered by speaker PN with its superimposed F0.

![Figure 1: An audio speech waveform (top) and a spectrogram (bottom) with a superimposed F0 trace of an example - [pʰaa mid-level tone] - uttered by speaker PN.](image)

From the F0 data points generated by *Praat*, the actual F0 used for the purpose of FVC were sampled at the 0%, 5%, 10%, 20%, 40%, 60%, 80%, 95%, 100% points of the phonetically voiced parts of the target segmental combinations. It should be pointed out here that the whole of the F0 time-course was used; not just the F0 on the Rhyme. This is because one wants to be able to make use of speaker-dependent information from the entire F0 sample.

2.3 Polynomial fitting of F0

In order to apply statistical techniques, we need to parameterize the variation of the F0 as it changes over the duration of the voiced part of the segmental combinations being spoken. To do this, we approximate the F0 data using a cubic polynomial function of type \( ax^3 + bx^2 + cx + d \), where \( a, b, c, \) and \( d \) are the coefficients and \( x \) is the time. The curve is defined by the value of the coefficients which we can use as parameters to describe the F0 contours. The coefficient values and curve fitting were done using scripts written in the R programming language (written by Dr Phil Rose, ANU).
Following is an example of a polynomial curve fitting in which the F0 values of the segmental sequence [paa] (high-falling tone uttered by speaker PN) are plotted together with its polynomial fitting.

![Polynomial fitting](image)

**Figure 2:** The sampled F0 values of the segmental sequence [paa] with high-falling tone uttered by speaker PN, plotted along time, together with a polynomial curve fitted.

In this study, we investigated the effect of the use of different discriminatory parameters. These different parameters are as follows:

1. **AllCoeff**: all coefficients (b, c and d in $ax^3 + bx^2 + cx + d$) of the cubic polynomial approximating the F0 shape.
2. **DurCoeff**: the duration, and the coefficient (d in $ax^3 + bx^2 + cx + d$) of the linear term of the cubic polynomial.
3. **AllVar**: all coefficients of the cubic polynomial describing the F0, and the duration.

3. **Experimental Results and discussion**

According to the experimental results involving the three parameters: **AllCoeff**, **DurCoeff**, and **AllVar**, most segmental combinations and tones, using all variables (**AllVar**) yielded the lowest EER% values. The average EER values of **AllVar** are lower than 10% for all tones except the low-rising tone (LH) of which the average EER is 15.3. On the other hand, the average EER% values are higher than 12% for the other two parameter settings (**AllCoeff** and **DurCoeff**) which are not presented here in this paper. Table 3 shows a summary of the forensic voice comparison tests (**AllVar** only) in terms of EER%.
Table 3: A summary of the forensic voice comparison tests (AllVar only) in terms of EER%.

<table>
<thead>
<tr>
<th>Segmental Combinations</th>
<th>M</th>
<th>L</th>
<th>HL</th>
<th>HH</th>
<th>LH</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AllVar</td>
<td>AllVar</td>
<td>AllVar</td>
<td>AllVar</td>
<td>AllVar</td>
<td>Ave.</td>
</tr>
<tr>
<td>muu</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>5</td>
<td>7.0</td>
</tr>
<tr>
<td>h'aa</td>
<td>14</td>
<td>5</td>
<td>20</td>
<td>18</td>
<td>18</td>
<td>15.0</td>
</tr>
<tr>
<td>buu</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>7.8</td>
</tr>
<tr>
<td>h'ii</td>
<td>12</td>
<td>13</td>
<td>11</td>
<td>13</td>
<td>18</td>
<td>13.4</td>
</tr>
<tr>
<td>bii</td>
<td>4</td>
<td>13</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>9.8</td>
</tr>
<tr>
<td>puu</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>4</td>
<td>26</td>
<td>11.0</td>
</tr>
<tr>
<td>h'uu</td>
<td>13</td>
<td>7</td>
<td>10</td>
<td>4</td>
<td>28</td>
<td>12.4</td>
</tr>
<tr>
<td>mii</td>
<td>10</td>
<td>14</td>
<td>10</td>
<td>7</td>
<td>20</td>
<td>12.2</td>
</tr>
<tr>
<td>pii</td>
<td>7</td>
<td>11</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>9.4</td>
</tr>
<tr>
<td>paa</td>
<td>13</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td>13</td>
<td>10.0</td>
</tr>
<tr>
<td>maa</td>
<td>7</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>8.0</td>
</tr>
<tr>
<td>baa</td>
<td>9</td>
<td>13</td>
<td>10</td>
<td>8</td>
<td>16</td>
<td>7.0</td>
</tr>
<tr>
<td>Ave.</td>
<td>9.1</td>
<td>9.8</td>
<td>9.8</td>
<td>9.1</td>
<td>15.3</td>
<td></td>
</tr>
</tbody>
</table>

In Table 3 the five contrastive Thai tones are phonetically realized as the mid-level tone (M), the low-level tone(L), the high-falling tone(HL), the high-rising tone(HH) and the low-rising tone(LH), respectively. AllVar are all variables (all coefficients and duration). The bold underlined figures are the best EER of each tone (in percentage).

It can be seen from Table 3 that the performance of the forensic voice comparison is very similar across the different tones (9.1% to 9.8% in terms of EER), except for the low-rising tone (LH) of which the EER is 15.3%. The poor performance in relation to the low-rising tone is interesting because it suggests that contour tones such as the high-falling and low-rising have a larger between-speaker variability, resulting in better discrimination performance. However, within-speaker variability may be equally large in those tones.

3.1 Tippet Plot Results
Since the EER is a binary metric, it does not indicate the different range of LRs that a given methodology produces. For example, a log LR value of 1 and a log LR value of 3 both supports the prosecution hypothesis, however, the strength of the support of the latter is two orders of magnitude greater than the other. This point
is crucial in presenting the evidence before the courts. Thus the result of each discrimination test is also assessed in terms of a Tippett plot.

Now we look at the results of the best performing segmental combinations using Tippett plots in Figure 3 (Owing to the space limitation, the worst performing segmental combinations are not shown in this paper). In the Tippett plots, the curves rising to the right (red curve) represent the cumulative proportion of the SS (same-speaker) comparisons with the log10LRs equal to or less than the value indicated on the x-axis while those rising to the left (blue curve) represent the cumulative proportion of the DS (different -speaker) comparisons with the log10LRs equal to or greater than the value indicated on the x-axis. Please note that the range of the x-axis is truncated from Log_{10}LR = 3.5 to Log_{10}LR = -10.5.

Thus it can be seen from Figure 3-1, for example, that about 4% of the SS comparisons (the y-axis value of the crossing point of the two curves) have log_{10}LR values which incorrectly suggest that they are more likely to come from different speakers (assuming we use the point where the SSLR and DSLR curves cross at a threshold).
From Figure 3, it is apparent that the SS comparisons produce much weaker log10LRs than the DS comparisons. For SS comparisons, the largest log10LR is between 1 and 2 whereas for DS comparisons, many of the Log10LRs are smaller than -10. In the low-rising tone of [muu] (see Figure 3-5), for example, about 55% of the DS comparisons have the Log10LRs higher than -10. According to the verbal equivalents of the likelihood ratios (Champod and Evett 2000: 240 in Rose 2002: 62), log10LR 1 to 2 indicates “moderate” evidence while log10LR > -4 indicates that the evidence is “very strong”. Thus, the SSLRs can only moderately support the prosecution's hypothesis, but many of the Log10LRs for DS comparisons can strongly support the defence hypothesis.
4. Conclusion

In this paper, the notion of forensic voice comparison (FVC) and its methodology using the likelihood ratio approach are presented. This is the first study to examine the discriminatory power of the tonal acoustic features of fundamental frequency (F0) and duration in Thai using Likelihood Ratio-based methodology.

The lowest EER obtained in this experiment was 4% (i.e. 96% of same-speaker speech data were correctly identified as coming from the same speaker). This suggests that FVC works reasonably well in Thai language. The results also show that FVC works best when the duration as well as the F0 coefficients are used as discriminatory variables. It has been demonstrated that the inclusion of duration in a parameter set significantly improves the performance of a forensic voice comparison. We also reported that as a whole, it is more difficult to identify individuals based on segmental combinations which are spoken in the low-rising tone in comparison to other tones. However, the results also showed that the performance of each tone depends on the selection of the particular segmental combinations.

It should be noted that the low EER% obtained in the experiments can be attributed to many favorable factors that may not be realistic in practice. The sound recording was conducted in well controlled conditions using high quality microphones in a sound proof room. Furthermore, the target segmental combinations were all recorded in a citation manner.

In sum, the result from this research has been promising and shows that it is beneficial to conduct further detailed research, such as using spontaneous speech samples and from both males and females with a larger number of speakers. As such, the next project aims to implement a wider study in discriminating speech in Thai. Various methods of modeling and discriminations will be trialed. Some standard methods of evaluation ($c_{th}$ and the 95% credible interval) will be employed to assess the performance of discrimination systems.
References


