F0 can tell us more: speaker verification using the long term distribution

Yuko Kinoshita\textsuperscript{1}, Shunichi Ishihara\textsuperscript{2}

\textsuperscript{1}Faculty of Arts and Design, University of Canberra, Australia
\textsuperscript{2}School of Culture, History and Language, the Australian National University, Australia

yuko.kinoshita@canberra.edu.au, shunichi.ishihara@anu.edu.au

Abstract

This study explores some options for improving the performance of F0-based speaker verification. We tested different parameterisation techniques that enable us to capture non-unimodal distribution. We also tested the use of dynamic features (delta F0) and different scales (log\textsubscript{10}). As a result, we discovered that combinations of these techniques could significantly improve both the performance of speaker verification, and the reliability of the likelihood ratios.

Index Terms: F0, forensic voice comparison, speaker verification, likelihood ratio

1. Introduction

Fundamental frequency (F0) is an attractive feature for forensic voice comparison. Although the mean and the standard deviation of F0 by themselves are ineffective in speaker verification \cite{1}, F0 was found to be a potentially useful addition when the shape of the long term F0 (LTF0) distribution was taken into account \cite{2}. \cite{2} parameterised the shape characteristics of LTF0 distribution by mean, standard deviation, skew, kurtosis, modal F0, and modal density. Using 201 male Japanese speakers, this experiment produced promising results with an equal error rate (EER) of 10.7\%. However, it also uncovered a weakness of this approach: many speakers’ LTF0 distributions are not unimodal, hence their characteristics cannot be captured by means of these six features. In fact, a bimodal distribution was found to be very common due to creaky phonation. Figure 1 shows examples of the LTF0 distribution presented in \cite{2}. The two lines (dotted and solid) show the LTF0 from two separate occasions.

![Figure 1: LTF0 distributions from two speakers.](image)

The six features proposed in \cite{2} do not capture the most striking feature of Speaker A – bimodality – at all. The figure for Speaker B shows a further problem. The distributions from two occasions have similar shapes for the second peak, but the strong first peak was observed only on one occasion. This first peak shifts the values of the six features significantly, and despite the general similarity of their shapes, these two distributions will produce very different values.

To improve the performance, this study explores options in two directions. Firstly we try out different techniques for modelling the shape of the distribution, which do not assume a unimodal distribution of the data. Secondly, we apply some commonly used front-end processing: logarithmic scale and dynamic features. We perform speaker verification tests with combinations of these different techniques, and study their effectiveness with respect to its discrimination performance and the calibration.

2. Approach

2.1. Recordings

For this study, we used 201 male Japanese speakers selected from the Corpus of Spontaneous Japanese (CSJ) \cite{3}. CSJ is a database which consists of various styles of speech recorded from 1464 speakers. The recordings used in this study were made in the style of either Academic Presentation Speech or Simulated Public Speech, and they were 12-25 minutes and 10-12 minutes respectively. 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 ‘spontaneity’ scale, in order to select speakers. By ‘spontaneous’, CSJ means ‘sounding as if it is not read out’. 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 other criterion for speaker selection was the availability of non-contemporaneous recordings. Our 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.

2.2. Testing features in this study

We attempted to improve the results of our previous study from two directions: 1) front-end processing of F0 prior to plotting the distribution, using 4 techniques, and 2) modelling of the shapes of the distribution, using 3 techniques. This made 12 different combinations of experimental conditions. We describe each step in detail below.

2.2.1. F0 extraction and processing

F0 in Hz (HzF0) was extracted using the ESPS routine of the Snack Sound Toolkit \cite{4} 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.

We used two scales for F0: a linear scale (Hz) and a logarithmic scale (log\textsubscript{10} Hz). We also included the dynamic feature, delta (Δ) F0 in this study, as previous research in automatic speech/speaker recognition reports that the addition of dynamic information improves performance (eg. \cite{5} \cite{6}). We obtained ΔHzF0 by simply taking the difference between two immediately adjacent HzF0 values (separated 0.05 second), i.e. ΔHzF0 = HzF0 – HzF0\textsubscript{i-1}. When one of these
two HzF0 values is $0$, its $\Delta$HzF0 was not calculated. The distributions of the delta values were then plotted. $\Delta \log_{10}$F0 was calculated from $\Delta$HzF0. Thus in terms of F0 processing we used both static (F0) and dynamic ($\Delta$ F0) features, and both linear (Hz) and non-linear scales ($\log\_10$). Thus there were 4 permutations tested: (HzF0, $\Delta$ HzF0, $\log\_10$F0 and $\Delta \log_{10}$F0).

2.2.2. Modelling of the distributions

The other set of variables was the techniques in modelling the shape of the distribution. Firstly, the probability density functions (PDF) of the sampled features were estimated for each speech sample, 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).

Then, three different techniques were applied to model the shape of the distribution. They were: 1) the six statistical features used in [2]; 2) 15% percentile; 3) 10% percentile.

Firstly we used the method in [2]; i.e. the values of the six features: long-term mean F0, standard deviation, skew, kurtosis, modal F0, and modal density. The results produced with this modelling technique with linear Hz F0 were used as the baseline features for the comparison in this study (‘baseline features’ hereafter).

For the other two options, we used percentiles. That is, referring to the x-axis (density) and y-axis (Hz or log10) values of a PDF at certain percentile points, we modelled the shape of the distribution. We used 15% and 10% percentile intervals in this study. That is, the x-axis and y-axis values of a PDF were extracted at the 5th, 20th, 35th, 50th, 65th, 80th and 95th percentiles for 15% interval (see Figure 2), and the 5th, 15th, 25th, 35th, 45th, 55th, 65th, 75th, 85th and 95th percentiles for 10% interval.

![Figure 2: An example PDF illustrating the modelling technique based on percentiles (15% interval).](image)

We expected the percentile techniques to characterise the shape of the distribution more accurately than the baseline features especially in cases like Figure 1, as it can capture multiple peaks. The numeric vectors extracted through these processes were used as the testing features for this study.

2.3. Speaker verification test

In the testing phase, each combination of speech samples was compared on the basis of the testing features described above. The likelihood ratio (LR) of observing the given difference under the competing hypotheses was estimated; and a speaker verification test was performed using the LR as a discriminant function. The details are described below.

2.3.1. Likelihood ratio (LR)

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 [7]. In the context of forensic voice comparison, it will be the probability of observing the difference between two speech samples if they had come from the same speaker (the ‘same speaker’ hypothesis) relative to the probability of observing the same evidence if it had been produced by different individuals (the ‘different speaker’ hypothesis). Letting $P$ represent probability, $E$ evidence, and $H$ hypothesis, LR can be expressed as (1):

$$LR = \frac{P(E|H)}{P(E|\bar{H})}$$

The LR will be larger than unity when the given evidence supports the same speaker’ hypothesis, and smaller than unity when the evidence support the different speaker’ hypothesis. The relative distance of the LR from unity quantifies the strength of the evidence.

2.3.2. LR calculation and testing in this study

In all discrimination experiments, two types of speaker pairs, non-contemporaneous same-speaker (SS) pairs and different-speaker (DS) pairs, were compared. With 201 speakers, 201 SS comparisons and 80400 DS comparisons are possible (201 speakers produced 20100 combinations of speakers, and each different-speaker pair produced four ways 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).

For the verification we therefore did not use a cross-validation approach, but assumed that the effect of the test data (2 speakers) not being independent of the population data (201 speakers) would be negligible in this case, because of the large number of speakers involved.

LRs for each comparison were estimated using the multivariate LR formula (MVLR) [8], which allows us to combine features that may be strongly correlated. This is an important consideration, as we expect many of the features used in this study to be correlated. The MVLR still has problems such as only accommodating two levels of variance, but a number of LR-based FSR discrimination experiments have shown that it is an effective technique to estimate LRs and classify speakers. (e.g. [9], [10], [11], [12], [13], [14]).

Once the LR was estimated, we proceeded to the verification test. If the given testing features are good predictors of the speakers’ identity, the LRs produced with them should be able to distinguish the same speaker pairs from different speaker pairs well, i.e. the SS pairs are expected to produce LRs greater than unity, whereas the DS pairs are expected to produce LRs smaller than unity. We assessed the verification performance with respect to how accurately it reported the correct same-speaker or different-speaker pairs.

2.3.3. Calibration of the LR

The error rates above show us how successfully we can make binary decisions on the speakers’ identity. However, in forensic voice comparison contexts, there is something else that analysts must consider and report: the calibration of the LR. For instance, in the context of forensic evidence presentation, although both LR 5 and 100 support the prosecution hypothesis, they have significantly different meanings, since LR 5 and 100 differ significantly in their strength of evidence. It is therefore extremely important to assess how well the LRs produced by a method are calibrated.
Thus, we examined the calibration of the LRs from the 12 experimental conditions using C_{lr} [15, 16].

3. Results and Discussion

3.1. Discrimination performance

Table 1 presents the error rate for the verification test with each of the 12 conditions. The error rates for SS comparisons and DS comparisons (the threshold set at unity) were also presented in order to observe the differences caused by the types of the speaker pairs. “Baseline” indicates the baseline features; 15% and 10% indicate the percentile values which were collected at interval of 15% and 10%, respectively. The rows “HzF0”, “Δ HzF0”, “log_{10}F0” and “Δ log_{10}F0” show the results based on the linear F0, delta F0, log_{10} F0 and delta of the log_{10} F0. In order to see the general tendency of each condition, the mean error rates were also presented. Table 1 also includes the equal error rate (EER), which is the point where the SS and DS comparison shares the same error rate.

Table 1 Error rates for each condition, presented separately for SS comparisons and DS comparisons.

<table>
<thead>
<tr>
<th></th>
<th>baseline</th>
<th>15%</th>
<th>10%</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>10.1</td>
<td>8.7</td>
<td>7.2</td>
<td>8.7</td>
</tr>
<tr>
<td>DS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HzF0</td>
<td>10.4</td>
<td>15.6</td>
<td>7.0</td>
<td>6.7</td>
</tr>
<tr>
<td>(EER)</td>
<td>13.4</td>
<td>7.8</td>
<td>4.6</td>
<td>5.6</td>
</tr>
<tr>
<td>Δ HzF0</td>
<td>7.5</td>
<td>8.8</td>
<td>7.0</td>
<td>7.3</td>
</tr>
<tr>
<td>(EER)</td>
<td>7.6</td>
<td>7.0</td>
<td>5.6</td>
<td>6.7</td>
</tr>
<tr>
<td>log_{10}F0</td>
<td>9.0</td>
<td>14.5</td>
<td>7.0</td>
<td>6.6</td>
</tr>
<tr>
<td>(EER)</td>
<td>11.4</td>
<td>6.6</td>
<td>4.1</td>
<td>7.3</td>
</tr>
<tr>
<td>mean</td>
<td>9.2</td>
<td>12.2</td>
<td>8.1</td>
<td>6.8</td>
</tr>
</tbody>
</table>

Figure 3 shows the differences in EER. It is clear that the general tendency of the performance is 10% percentile > baseline.

Figure 3 EERs produced with the 12 different experimental conditions

Firstly we observe the effect of the variables related to F0 processing. At first glance, it appears that F0 processing seems not to have a consistent effect across the three techniques used for distribution capturing. However, if we look at the results for the baseline and the two types of percentile separately, we can see some tendencies. In both 15% and 10% percentiles, HzF0 had the worst result, and Δ log_{10}F0 had the best. The results appear to suggest two points: 1) log_{10}F0 performs better than HzF0; 2) taking delta improves the performance. The inconsistency between baseline features and the two percentiles is not surprising, as they take very different approaches to quantifying the shape of the distribution.

Now turning to the effect of the difference in the techniques for modelling the distribution shape, we can see a clear tendency. As we predicted, both types of percentile seem to outperform the baseline features. In every condition but one (15% percentile/HzF0), the baseline was outperformed by both 15% and 10% percentile. This was consistent for both SS and DS comparisons. The comparison between 15% and 10% percentile shows that 10% seems to produce better results. When we used the distribution of log_{10}F0, the SS comparison had more or less the same result for 15% and 10% percentile but, in all other conditions, the 10% percentile produced better results, which probably relates to its finer sampling of the shape of the distributions.

In summary, we found that there are multiple ways to make improvements on the previous results. The use of a non-linear scale (log_{10}) and the inclusion of a dynamic feature (Δ F0) can contribute to the improvement. We can also improve the results by changing how we model the shape of the distribution. Both percentile 10% and 15% outperformed the baseline features used in previous study, and percentile 10% performed best among all three. It seems plausible that a still smaller percentile with more sampling points could improve the performance further, but this needs to be tested.

3.2. Calibration of the LR

As described above, C_{lr} is a metric that can assess two things: 1) how well the LRs produced by a system are calibrated, 2) how well the system makes binary decisions. Those two components in C_{lr} can be broken down to two separate scores, C_{lr,cal} and C_{lr,min}. C_{lr,cal} shows how much loss the system has in its calibration component, whereas C_{lr,min} shows how well the system can make binary decisions when the system is ideally calibrated. In both cases, the smaller the scores are, the better the system. Figure 4 presents the C_{lr} for each testing condition with the break up for C_{lr,cal} and C_{lr,min}.

Figure 4 C_{lr,cal} and C_{lr,min} for each condition

It appears that taking delta improves C_{lr}, especially C_{lr,cal}, and this difference between the delta and the non-delta feature sets in C_{lr,cal} is found to be statistically significant (p = 0.006). Taking log_{10} did not have a consistent effect.

The observation of C_{lr,min} shows that the distribution modelling techniques have a very consistent effect. In all cases, 10% percentile produced the best score, then percentile 15%, and the baseline features were the worst. The degree of the improvement was greater for dynamic features than static features. This result is consistent with what we observed with EER, understandably, since both are the indicators of how well the system in question makes binary decisions.

Figure 5 shows the Tippett plots of the baseline and the best performing condition (Δ log_{10}F0/10% percentile). The red and black lines for the SS and DS comparisons, and the
dotted and solid lines are for pre- and post-calibration LRs, respectively. They are very similar with respect to their shapes, but the latter has lower EER (4.1%), indicating it had the better verification performance.

![Figure 5 The Tippett plots of the baseline and the best performing condition (Δlog₁₀F₀; 10% percentile).](image)

Table 2 presents the results of the verification based on the calibrated LRs ([15] [16]). Although the calibration improved the performance of SS comparisons, that for DS comparisons deteriorated. This reflects that the calibration shifted the whole distribution to the right along the x-axis, which is consistent with various previous studies (e.g. [9] [10]). In other words, the calibration made the reliability of these two types of verification more comparable by reducing the difference in their error rates.

<table>
<thead>
<tr>
<th>Table 2 Post-calibration verification results</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>HzF₀</td>
</tr>
<tr>
<td>ΔHzF₀</td>
</tr>
<tr>
<td>log₁₀F₀</td>
</tr>
<tr>
<td>Δlog₁₀F₀</td>
</tr>
</tbody>
</table>

4. Conclusion
This study revealed that speaker verification based on LTF₀ as reported in [2] can be improved considerably by using a dynamic feature and a non-linear scale rather than the measurement based on static F₀ in a linear scale, and by capturing the shape of the distribution based on percentiles. This percentile-based technique for capturing a distribution had a very consistent effect, and 10% percentile was found to be most effective. With the front-end F₀ processing, the results varied depending on whether the distribution shape was captured by the baseline features or by percentile. Using percentile, which was found to be more effective, the use of a dynamic feature and a perceptual scale improved the verification performance.

In the observation of the calibration, we revealed that the use of the dynamic feature has a significant positive effect on the quality of LRs produced. Cᵣₐ was improved significantly, which implies the LRs produced were much more reliable. In the context of forensic voice comparison, this is an important benefit.

5. Acknowledgements
We would like to thank our anonymous reviewers for their valuable comments. This research is financially supported by the College of Asia and the Pacific, ANU.

6. References