Speech Length Threshold in Forensic Speaker Comparison by Using Long-Term Cumulative Formant (LTCF) Analysis

CAO Honglin$^{1,2}$, KONG Jiangping$^2$

$^1$ Key Laboratory of Evidence Science (China University of Political Science and Law), Ministry of Education, China;
$^2$ Dept. of Chinese Language and Literature, Peking University, Beijing, China;
caohonglin@pku.edu.cn; jpkong@pku.edu.cn

ABSTRACT: Long-Term Formant distribution (LTF) is a relatively new method in forensic speaker comparison, by which the results have been proved to contain important speaker-specific information. However, few studies have been carried out for the fundamental issue that how long the speech sample should be collected. The current paper investigated the speech length threshold (SLT) by using Long-Term Cumulative Formants (LTCF) analysis, which was one of the LTF methods. The speech sample for each speaker was segmented into one-second length subsamples. Pearson’s correlation coefficients were calculated for LTCF values of the whole speech sample and new set of speech samples that were formed by adding the immediately following subsample onto the speech sample before it with a start from the first subsample. The results show that SLT can be placed at about 70 seconds natural speech recordings (approximate 20 seconds only vocalic samples in duration), which are adequate to represent the whole vocal tract resonance characteristics.

Keywords: long-term formant; speech length; correlation coefficients; forensic speaker comparison

1. INTRODUCTION

Formant features are of great importance in forensic speaker identification, as they contain lots of vital speaker-specific information. Recently, Nolan and Grigoras [1], in a case study, proposed a new method of formant analysis, which was called the Long-Term Formant distribution (LTF). Instead of selecting specific vowel targets, this method captures the information from all vocalic portions, from which formant structures are visible and reliable, leading to a long-term distribution for each formant. It can summarize the whole resonance characteristics of individual vocal tract and reflect an individual’s anatomy and articulatory habits [2]. Many advantages of the LTF method are found, including relatively time-efficient application, high inter-expert reliability, anatomical motivation (LTF2 and LTF3 are negatively correlated with speaker height) and language independence [3, 4]. Meanwhile, the LTF values (and the bandwidths) of F1 to F3 can also be applied to automatic speaker recognition effectively [5].

As for any long-term feature (e.g. Long-Term Average Spectrum and Long Term F0 distribution) in forensic phonetics, the speech length threshold (SLT) is an important influencing factor. In previous studies, speeches in varied lengths were used. For instance, in [6, 7], both spontaneous and reading materials were adopted. The durations of the two style recordings were, on average, 178s (range from 79s to 313s ) and 39s (range from 31s to 54s) respectively. After editing for LTF analysis (i.e. only vocalic portions with clear formant structure remained), the durations decreased correspondingly, and the values were, on average, 40s (range from 12s to 83s) and 12s (range from 8 to 16s) for the two style recordings respectively. In [5], 22s and 11s length of samples (after editing) were investigated for training and test set respectively. In [2], 30s of vowel/approximant materials were selected for LTF analysis. Are these durations above long enough for LTF analysis? According to [8], it is recommended that, using the Catalina (Version 3.0h) software, speech (raw materials without editing) lengths should be longer than 10 seconds. Additionally, Moos [6, 9] compared the standard deviation (SD) of LTFs (F1 to F3) of different packages (durations were cumulative from 1s to 10s with an interval of 0.5s). The author found that there was a net duration threshold value of available speech material beyond which LTF’s were saturated, which could be placed at around 5s to 8s of pure vocalic stream, depending on the formant and the speaking condition. About 6 seconds of pure vocalic stream (equivalent to 27 seconds of dialogue or 19 seconds of read speech) were, on average, enough to produce reliable LTF values [7].

Most studies about LTF analysis were based on English and German speech and focused on Long Term Average Formants (LTAF). Actually, there are two types of LTF analysis including LTAF and Long Term Cumulative Formants (LTCF) [8]. LTAF shows the distribution of each formant individually, while LTCF represents the
vertical addition of all LTAFs (cf. Fig. 1). The present study, based on Chinese reading materials, examines the correlations of LTCFs of speech with a variety of durations, in order to test the discrimination ability and SLT by using LTCF analysis.

2. METHOD

2.1. Material

Six short passages were selected in this study. They are short introductions of six famous cities in China: Beijing, Shanghai, Shenzhen, Hong Kong, Xi'an and Guangzhou (chosen from [10]). All of texts contain 2,919 syllables (In Standard Chinese one character stands for one syllable [11]). The length of the speech materials is about 10 to 20 times the length of general studies’ materials, which can be shown from Table I (cf. the duration information of the references described above). The reason for this is to guarantee the hypothesis that this speech length is adequate to represent the speaker's LTCF distribution characteristics is reasonable.

2.2. Speakers and Recording

The subjects included three male native speakers of Standard Chinese, aged 30 to 34, none of whom had any noticeable voice and speech disorders (owing to the data processing process was very time-consuming, only three speakers were investigated eventually). They were required to be familiar with the materials first and then read the text at a normal speed, in a comfortable way. The SONY ECM-44B microphone was used to record the materials in a sound attenuated room at Peking University. All recordings were made at a sampling rate of 22 kHz and 16 bit depth.

2.3. Data Processing

First of all, a short C program using dynamic programming algorithm was carried out on eliminating all voiceless information. Then Wavesurfer [12] was chosen to edit and analyze the materials. The recordings were edited by hand to eliminate all nasal consonants and other vocalic portions where the formant structures were unclear. The durations of the original recordings and resulting samples of the three speakers were shown in Table I. Subsequently, formant tracking was applied to all resulting samples by using the LPC-based algorithm. The first four formant values were obtained automatically, checked and, if necessary, corrected manually. These procedures were all applied with Wavesurfer, with the settings as follows: LPC order : 12, number of formants : 4, analysis window length (hamming): 0.049s, pre-emphasis factor: 0.7, frame interval: 0.01s, down-sampling frequency: 10 kHz.

Table 1. The durations of the original recordings and resulting samples of the three speakers. “m” and “s” stand for minute and second respectively.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Original duration</th>
<th>Resulting duration</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH</td>
<td>11m30s</td>
<td>3m21s (201s)</td>
<td>29%</td>
</tr>
<tr>
<td>YQ</td>
<td>12m57s</td>
<td>3m42s (222s)</td>
<td>29%</td>
</tr>
<tr>
<td>YF</td>
<td>17m26s</td>
<td>4m52s (292s)</td>
<td>28%</td>
</tr>
</tbody>
</table>

A MatLab program was used to extract formant values, generate histogram, display the distribution of values and calculate the correlation coefficients. As discussed above, many studies focused on the mean or mode values of LTAF (often F1 and F3). However, the shape, like kurtosis and skewness, of the distribution of each formant is also very useful, because the LTAF distributions are often non-symmetrical. In the present study, we use the LTCF values for calculating, other than the mean (or mode) value of each LTAF. The differences of LTCF and LTAF are shown in Fig. 1. The LTCF represents the vertical addition of all LTAFs, i.e. the LTCF shows the cumulative frequency of occurrence of all values of F1 to F4. For one LTCF, it can be represented by an array, which is composed of the number of the frequency of occurrence (Y-axis), with a fixed X-axis range. The array of one LTCF is easily-acquired and conveniently-calculated. Meanwhile, it includes the shape information, even though it doesn’t differentiate the detailed information of each LTAF distribution.

Figure 1. An illustration of LTAF and LTCF of Speaker YQ. All histograms are displayed in 25 Hz bins.

It is hypothesized that the whole resulting speech samples, whose lengths are shown in Table I, are adequate enough
to summarize the resonance characteristics of the three speakers’ vocal tracts. The whole speech sample for each speaker is segmented into one-second length subsample. New set of speech samples (called cumulative samples) that are formed by adding the immediately following subsample onto the speech sample before it with a start from the first subsample are studied. For each speaker, the LTCF values of cumulative samples are calculated. For example, the length of 1s, 2s…200s and 201s cumulative samples of speaker CH are analyzed. And then, Pearson’s correlation coefficients between the LTCF values of the whole speech sample and the LTCF values of the cumulative samples are calculated. To ensure the LTCF values of different samples can be calculated, the range of the whole formants frequencies is normalized to 200-4500Hz, which can cover the extreme of the F1-F4 values of the three speakers by the square.

3. RESULTS

Pearson’s correlations (r-values) between the LTCF values of every cumulative sample and the LTCF values of the whole speech samples of three speakers are shown in Fig. 2 in the form of line chart. The red, blue and green curves represent speaker CH, YQ and YF respectively. It shows , for example, that the r-value (for all r-values, p>0.000) of speaker CH increases from 0.65 to 0.90 rapidly, when the duration is between 1s to 9s. From 10s to about 28s of the cumulative duration, the r-value increases from 0.90 to 0.96 on medium speed. In the range of 29 to 103s, the change of r-value is relatively stable (the r-values range from 0.96 to 0.97). From 104s to the end of the material, another increase appears. To get a better view of the duration differences between the three speakers in terms of certain r-values, some special duration information is also displayed in Table II, when the r-values reach 0.80, 0.85, 0.90 and 0.95. There are at least two findings that can be derived from Fig. 2 and Table II. The first important one is that all three speakers’ r-values move up at a decreasingly rapid rate, as the duration increases. The second finding concerns the individual differences of the three speakers: the distributions of the three curves are not exactly the same. For instance, when the duration is 5s, the r-value of speaker YQ is 0.80, but the value that speaker YF can reach is 0.85 instead; when the r-value reaches 0.95, no less than 18s and 24s long vocalic samples are needed for speaker YQ and YF respectively; it seems that less vocalic samples are needed for speaker YQ than speaker CH to reach a certain r-value.

Table 2. The duration information (in second) of three speakers, when r-values reach 0.80, 0.85, 0.90 and 0.95.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>r = 0.80</th>
<th>r = 0.85</th>
<th>r = 0.90</th>
<th>r = 0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>YQ</td>
<td>5</td>
<td>6</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>YF</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>mean</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>21</td>
</tr>
</tbody>
</table>

Fig. 3 shows the mean r-values distribution of three speakers. Specifically, for each speaker, r-values between the LTCF values of every cumulative sample and the LTCF values of the whole speech samples of the same speaker (intra-speaker correlations) are calculated. The red curve represents the mean r-values of the three curves which are shown in Fig. 2. Additionally, for each speaker, by contrast, the same LTCF values of cumulative samples are also compared to the other two speakers’ whole speech samples’ LTCF values (inter-speaker correlations), and then 2*3=6 groups of r-values are gotten. The blue curve represents the mean values of the 6 groups. The magenta and green curves represent the distributions of the mean values plus and minus the SD of the 6 groups respectively. Markedly, the red curve is much “higher” than the other three curves, which means that it is easy to discriminate the three different speakers using the correlation coefficients calculated by LTCF method. Meanwhile, the discrimination ability of correlation coefficients of LTCFs can be significantly improved, as the cumulative durations increase. It can be seen that, approximately, 20 seconds in duration is the boundary between the stable and non-stable change of the discrimination ability.

To get a better view of the change rate of r-values as the duration increases, Fig. 4 is generated. In Fig. 4, the red
and blue curves represent the SD values of every three consecutive points (along the X-axis) of the red and blue lines in Fig. 3 respectively. The range of the X-axis is limited to 1s to 100s so that the two curves can be more clearly illustrated. As can be observed from Fig. 3-4, the r-values are volatile when the length of the vocalic sample is less than about 20s. The SDs of the average r-values of both intra- and inter-speaker correlations are under 0.0025 constantly when the duration is longer than about 20s.

![Figure 3](image)

**Figure 3.** The mean r-values distribution of three speakers. The range of the X-axis is limited to 1 to 200 seconds.

![Figure 4](image)

**Figure 4.** The change rate of r-values. The red and blue curves represent the SD of the three consecutive points of the red and blue lines in figure 3, respectively.

### 4. DISCUSSION AND CONCLUSION

In this paper, about 20s vocalic samples in duration are found to be able to have a stable discrimination ability of the three speakers by using LTCF analysis, i.e. the LTCF values of 20s vocalic samples seem to be able to show a good representation of the whole resonance characteristics of individual vocal tract. Table I shows the proportional relationship between the duration of resulting vocalic samples and original recordings, which is 29% on average (cf. [7]). Based on this proportion, about 70s original speech (speaking at natural speed) in length is needed, which is equivalent to 20s vocalic samples. From a forensic point of view, 70s speech in length of one speaker is often available in civil cases and also in a few criminal cases. It doesn’t mean that, however, the speech length less than 70s (or 20s vocalic samples) cannot be used. The results presented in this paper show that the r-values of all three speakers can reach 0.90 when 9s vocalic samples (about 31s natural speech) are available. In a small sample size (number of speakers), 9s will be adequate to discriminate speakers. But as the sample size increases, more speech materials will be needed, since the possibility that the r-value of two speakers’ LTCFs is very high (e.g. exceed 0.90) cannot be ruled out. Meanwhile, the relation of r-value and duration varies between the three speakers, which supports Moos’ findings [7, 9].

Many previous studies on LTF analysis focused on the mean or mode value of every LTAF (often F1 and F3) based on English or German speech materials. The approach proposed in this paper uses LTCF values based on Chinese reading materials instead. So the results of these studies are hard to be compared directly. To some extent, for different languages, it can be speculated that 20s vocalic speech should be adequate for LTF analysis (both LTAF and LTCF). Anyhow, the longer the better. It is worth mentioning that another direction for future research will be calculating the relation of each single LTAF other than the whole LTCF, because it was found that the discrimination abilities of LTAFs of different formants were not the same (e.g. Moos [6, 7, 9] found that LTF value of F3 was most valuable). It will be useful not only for the quantitative analysis on LTAF distribution of every formant, but also for the new understanding of SLT by using LTF analysis.

In conclusion, the present paper proposed a new method to discriminate speakers using LTCF analysis based on three male speakers’ long Chinese reading materials. For each speaker, Pearson’s correlation coefficients between the LTCF values of cumulative samples and the LTCF values of the whole speech sample are calculated. The results show that this method can distinguish the three speakers effectively and the more speech materials are available, the steadier the dis crimination ability is. Markedly, on average, the speech length threshold can be placed at about 70 seconds natural speech recordings (approximate 20 seconds only vocalic samples in
duration), which are adequate to represent the whole resonance characteristics of individual vocal tract and are enough to produce reliable LTCF values. The findings reported here also have significant theoretical value to forensic casework. However, owing to only three subjects and reading other than spontaneous speech were studied, this research has some limitations. In the future studies, the effects of sample size (number of speakers), different speaking styles, languages and contents of the speech materials still have to be further investigated, which are very important factors in forensic applications.

5. ACKNOWLEDGMENTS

This research is funded by the Ministry of Sciences and Technology of China. Grant No: 2009BAH41B00. Many thanks to Dr. Michael Jessen for introducing the LTF method used in BKA. Thanks to Dr. Zhu Lei for writing the C program to edit the original materials. We also thank Dr. Lee Yinghao and Zhang Wei for their comments for the early draft of the article.

6. REFERENCES