HIGH IMPROVEMENT OF SPEAKER IDENTIFICATION AND VERIFICATION BY COMBINING MFCC AND PHASE INFORMATION

Longbiao Wang¹, Shinji Ohtsuka², Seiichi Nakagawa²

¹Department of Systems Engineering, Shizuoka University, Japan
²Department of Information and Computer Sciences, Toyohashi University of Technology, Japan

wang@sys.eng.shizuoka.ac.jp, {ohtsuka, nakagawa}@slp.ics.tut.ac.jp

ABSTRACT

In conventional speaker recognition methods based on MFCC, phase information has been ignored. We proposed a method that integrated the phase information with MFCC on a speaker identification method, and a preliminary experiment was performed. In this paper, we propose a new method, and a preliminary experiment was performed. In this paper, speaker verification experiments were evaluated using speech database consisting of normal, fast and slow speaking rates. The speaker identification error rate was remarkably reduced for normal, fast and slow speaking rates in comparison with a standard MFCC-based method. In this paper, speaker verification experiments were also evaluated using the phase information. The experiments show that the phase information is also very useful for the speaker verification.

Index Terms: speaker identification, speaker verification, MFCC, phase information, combination method

1. INTRODUCTION

For text-dependent speaker recognition, different types of speaker models have been studied. Hidden Markov models (HMM) have become the most popular statistical tool for this task. The best results have been obtained using continuous density HMM (CHMM) for modeling the speaker characteristics. For the text-independent task, the temporal sequence modeling capability of the HMM is not required. Therefore, one state CHMM, also called a Gaussian mixture model (GMM), has been widely used as a speaker model. The use of GMM for modeling speaker identity is motivated by the fact that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities. Further, we proposed a novel method by combining speaker-specific GMM and speaker-adapted HMM [1].

Several studies have indicated a large effort to directly model and incorporate the phase into the recognition process [2, 3]. The importance of phase in human speech recognition has been reported in [4, 5]. Especially, the phase may be important for speaker recognition, because it may convey the source information. However, in conventional speaker recognition methods based on MFCC, it only utilizes the magnitude of the Fourier Transform of the time-domain speech frames. This means that the phase component is ignored. The MFCC captures the speaker-specific vocal tract information. Feature parameters extracted from excitation source characteristics are also useful for speaker recognition [6, 7, 8, 9]. We proposed a speaker identification method using phase information which was integrated with MFCC-based GMM [10] by using speech database consisting of normal speaking modes. In this paper, we propose an improved method and evaluate it by using speech database consisting of normal, fast and slow speaking modes. To improve the speaker identification performance, the MFCC-based GMM is expanded to a combination of MFCC-based GMM and MFCC-based HMM, and it is integrated with both the original and new phase information.

Speaker verification [11, 12] is the other important issue of speaker recognition which is even more successful in commercial systems than speaker identification. The speaker verification task is to decide whether or not an unlabelled voice sample belongs to a specific reference speaker. For GMM-based speaker verification, the likelihood of claimed speaker model given the speech segment is used. Therefore, we expect that the phase information is also effective for speaker verification. In this paper, the new phase information is used to perform speaker verification using the same experimental setup for speaker identification.

2. PHASE INFORMATION ANALYSIS

2.1. Formula

The spectrum $S(\omega, t)$ of a signal is obtained by DFT of an input speech signal sequence

$$S(\omega, t) = X(\omega, t) + jY(\omega, t) = \sqrt{X^2(\omega, t) + Y^2(\omega, t)} \times e^{j\theta(\omega, t)},$$

where in the other frequency $\omega'$ changes depending on the clipping position of the input speech even with a same frequency $\omega$. To overcome this problem, the phase of a certain basis frequency $\omega$ is kept constant, and the phase of other frequency is estimated relatively. For example, setting the basis frequency $\omega$ to $\pi/4$, we have

$$S'(\omega', t) = \sqrt{X^2(\omega', t) + Y^2(\omega', t)} \times e^{j\theta(\omega', t)} \times e^{j\theta'(-\omega, t)},$$

where in the other frequency $\omega'$ = $2\pi f'$, the spectrum becomes

$$S'(\omega', t) = \sqrt{X^2(\omega', t) + Y^2(\omega', t)} \times e^{j\theta(\omega', t)} \times e^{j\theta'(-\omega, t)}$$

$$= (X(\omega', t) + jY(\omega', t)),$$
with this, the phase can be normalized. Then, the real and imaginary part of Equation (3) becomes

\[
\tilde{X}(\omega', t) = \sqrt{X^2(\omega', t) + \tilde{Y}^2(\omega', t)} \times \cos\{\theta(\omega', t) + \frac{\omega'}{\omega} (\frac{\pi}{4} - \theta(\omega, t))\},
\]

(4)

\[
\tilde{Y}(\omega', t) = \sqrt{X^2(\omega', t) + \tilde{Y}^2(\omega', t)} \times \sin\{\theta(\omega', t) + \frac{\omega'}{\omega} (\frac{\pi}{4} - \theta(\omega, t))\}.
\]

(5)

In the experiments of this paper, the basis frequency \(\omega\) is set to \(2\pi \times 1000\ Hz\). In a previous study [10], to reduce the number of features parameters, we used only phase information in a sub-band frequency range. However, it was a problem for this method when comparing two values of phase. For example, when the two values are \(\pi - \theta_1\) and \(-\pi + \theta_1\), the difference becomes \(2\pi - 2\theta_1\). If \(\theta_1 \approx 0\), then the difference \(\approx 2\pi\), nevertheless the two phases are very similar each other. Therefore, in this paper, we modified the phase into coordinates on a unit circle, that is,

\[
\theta \rightarrow \{\cos \theta, \sin \theta\}
\]

(6)

### 2.2. Examples

We generated speech wave by using speech synthesis simulator "VT-Cales" [13], which can control the vocal source wave, pitch (F0) and vocal tract shapes. Fig. 1 illustrates the normalized phase and power spectrum by various conditions. We fixed the vocal tract shape which corresponds to vowel /a/. As shown in Fig. 1, the phase is much more influenced by vocal source characteristics, that is, speaker characteristics. Of course, the phase is also influenced by the vocal tract shape. Therefore the distribution of phase for a speaker should be modelled by mixtures of Gaussion.

### 3. COMBINATION METHOD AND DECISION METHOD

In this paper, the GMM (or HMM) based on MFCC is combined with the GMM based on phase information. When a combination of two methods is used to identify/verify the speaker, the likelihood of MODEL 1 is lineary coupled with that of MODEL 2 to form a new score \(L_3^n\) given by

\[
L_3^n = (1 - \alpha)L_{MODEL1}^n + \alpha L_{MODEL2}^n
\]

(7)

for the combination of three models, the new score \(L_3^n\) is given by

\[
L_3^n = (1 - \beta)L_{MODEL1}^n + \beta ((1 - \alpha)L_{MODEL2}^n + \alpha L_{MODEL3}^n)
\]

(8)

where \(L_{MODEL}\) is a likelihood produced by the \(n\)-th speaker MODEL, \(n = 1, 2, \ldots, N\), where \(N\) is the number of speakers registered. \(\alpha\) and \(\beta\) denote weighting coefficients, respectively.

For speaker identification, a speaker with maximum likelihood is decided as the target speaker. For speaker verification, it is to decide whether or not an unlabelled voice sample belongs to a specific reference speaker according to the likelihood ratio. Therefore, the likelihood normalization is a very important issue to deal with real-world data for speaker verification. Cohort-based normalization [14] is one of the effective normalization method for speaker verification. The cohort-based normalization uses a set of cohort speakers who are close to the target speaker, which is very easy to implement and could obtain high performance. Therefore, the cohort-based normalization is used in this paper. The size of cohort was set to 3.

### 4. EXPERIMENTS

#### 4.1. Database and Speech Analysis

We used the NTT database for the experiments. The NTT database consists of recordings of 35 speakers (22 males and 13 females) collected in 5 sessions over 10 months (1990.8, 1990.9, 1990.12, 1991.3 and 1991.6) in a sound proof room [15]. For training the models, 5 same sentences for all speakers from one session (1990.8) were used. They were uttered by a normal speaking style mode. Five other sentences every the four sessions were uttered at normal, fast and
slow speeds and used as test data. That is to say, the test corpus consisted of 2100 trials for speaker identification, and 2100 true trials and 71400 false trials for speaker verification. The average duration of the sentences is about 4 seconds. The input speech was sampled at 16 kHz. 12 MFCCs were calculated at every 10 ms with a window of 25 ms. The spectrum with 128 components consisting of magnitude and phase was calculated by DFT for every 256 samples. For phase information, the first 12 feature parameters, that is, from the 1st component to 12th component of the spectrum (frequency range: 60 Hz - 700 Hz) which obtained the best identification performance among all other sub-band frequency range [10] were used.

### 4.2. Speaker Identification Results

We evaluated the speaker identification experiment using phase information. GMMs with 64 mixtures having diagonal covariance matrices were used as speech models. The speaker identification results by individual method and combination method are shown in Table 1. The results with underline in Table 1 are the results of the preliminary experiment in our previous study [10].

Table 1. Speaker identification result (%)

<table>
<thead>
<tr>
<th>speed</th>
<th>normal</th>
<th>fast</th>
<th>slow</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC_GMM</td>
<td>97.7</td>
<td>95.1</td>
<td>94.9</td>
<td>95.9</td>
</tr>
<tr>
<td>{θ}</td>
<td>52.6</td>
<td>51.6</td>
<td>51.7</td>
<td>52.0</td>
</tr>
<tr>
<td>{cos θ, sin θ}</td>
<td>73.4</td>
<td>72.0</td>
<td>70.4</td>
<td>71.9</td>
</tr>
<tr>
<td>MFCC_GMM+{θ}</td>
<td>99.0</td>
<td>97.0</td>
<td>96.6</td>
<td>97.5</td>
</tr>
<tr>
<td>MFCC_GMM+{cos θ, sin θ}</td>
<td>99.3</td>
<td>98.0</td>
<td>98.1</td>
<td>98.5</td>
</tr>
<tr>
<td>MFCC_GMM+MFCC_HMM</td>
<td>99.3</td>
<td>97.9</td>
<td>96.6</td>
<td>97.9</td>
</tr>
<tr>
<td>MFCC_GMM+MFCC_HMM+{θ}</td>
<td>99.4</td>
<td>98.6</td>
<td>97.4</td>
<td>98.5</td>
</tr>
<tr>
<td>MFCC_GMM+MFCC_HMM+{cos θ, sin θ}</td>
<td>99.4</td>
<td>98.9</td>
<td>98.9</td>
<td>99.1</td>
</tr>
</tbody>
</table>

The speaker modeling techniques based on GMM and HMM which are used for speaker identification are also used for speaker verification [11, 12, 14]. The experimental setup of MFCC-based GMM and MFCC-based HMM is same as that of speaker identi- 

<table>
<thead>
<tr>
<th>speed</th>
<th>normal</th>
<th>fast</th>
<th>slow</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC_GMM</td>
<td>0.58</td>
<td>1.28</td>
<td>1.38</td>
<td>1.08</td>
</tr>
<tr>
<td>MFCC_HMM</td>
<td>0.70</td>
<td>1.00</td>
<td>1.43</td>
<td>1.05</td>
</tr>
<tr>
<td>{cos θ, sin θ}</td>
<td>5.22</td>
<td>6.15</td>
<td>5.69</td>
<td>5.69</td>
</tr>
<tr>
<td>MFCC_GMM+MFCC_HMM</td>
<td>0.35</td>
<td>0.91</td>
<td>1.16</td>
<td>0.81</td>
</tr>
<tr>
<td>MFCC_GMM+{cos θ, sin θ}</td>
<td>0.38</td>
<td>0.75</td>
<td>0.77</td>
<td>0.63</td>
</tr>
<tr>
<td>MFCC_HMM+{cos θ, sin θ}</td>
<td>0.28</td>
<td>0.40</td>
<td>0.83</td>
<td>0.50</td>
</tr>
<tr>
<td>MFCC_GMM+MFCC_HMM+{cos θ, sin θ}</td>
<td>0.18</td>
<td>0.37</td>
<td>0.71</td>
<td>0.42</td>
</tr>
</tbody>
</table>

4.3. Speaker Verification Results

The effectiveness of use of the phase information on speaker identification was demonstrated in Section 4.2. However, the number of registered speakers was not very large. To evidence the robustness of phase information for speaker recognition, the new phase information \{cos θ, sin θ\} is used to perform speaker verification in this section.

The Equal Error Rate (EER) of speaker verification is shown in Table 2. The trend of speaker verification result is similar to speaker identification result. Although the new phase information

---

1. GMMs with 32 mixtures were also used for speaker identification, however, it worked worse than GMMs with 64 mixtures. Due to limited space, the result based on GMMs with 32 mixtures was not described in this paper.

2. The result is slightly different from that in [10] because the experimental setup is a little different.

3. For speaker verification, the performance of the original phase information \{θ\} was significantly worse than that of the new phase information \{cos θ, sin θ\}, it was not described in this paper due to limited space.

4. In fact, the population size is a critical performance parameter for speaker identification, with the probability of error approaching 1 for indefinitely large populations. However, the performance of speaker verification is unaffected by the population size [18].
MFCC_GMM
MFCC_HMM
MFCC_GMM+Phase_GMM
MFCC_GMM+MFCC_HMM
MFCC_GMM+Phase_GMM

Section Rate (%)
Phase_GMM MFCC_GMM
MFCC_HMM MFCC_GMM+MFCC_HMM
MFCC_GMM+Phase_GMM

The Detection Error Trade-off (DET) curves of different feature parameters by DET curves

\( \{\cos \theta, \sin \theta\} \) worked worse than MFCC, a relatively high verification performance (about 5% ERR) was obtained. Comparing to speaker identification, the complement of MFCC-based GMM and MFCC-based HMM is relatively small. Because of the high complement of the MFCC (vocal tract information) and the phase information (vocal source information), the combination of MFCC and phase improved the speaker verification performance remarkably. The combination of the MFCC-based GMM and the phase based GMM achieved a relative error reduction rate of 40.4% from the MFCC-based GMM, and the combination of the MFCC-based HMM and the phase based GMM achieved a relative error reduction rate of 52.4% from the MFCC-based HMM. When the three methods were integrated, a relative error reduction rate of 61.5% from MFCC-based GMM, 60.0% from MFCC-based HMM and 49.4% from the combination of MFCC-based GMM and MFCC-based HMM were achieved. The Detection Error Trade-off (DET) curves of different feature parameters were compared in Fig. 2. Using the phase information, the trade-off of false acceptance rate and false rejection rate is much better than that based on MFCC only.

5. CONCLUSION

We proposed a text-independent speaker recognition method by combining MFCC and newly defined phase information. The speaker identification experiments were conducted on NTT database which consists of sentences data uttered at normal/slow/fast speed mode by 35 Japanese speakers. The proposed new phase information \( \{\cos \theta, \sin \theta\} \) remarkably improved the identification performance from the original phase information \( \theta \) for all speaking modes. Combining the MFCC and phase information, we obtained the error reduction rate of 52.2%, 55.6% and 59.2% than MFCC for normal, slow and fast speaking modes, respectively. Combining the MFCC-based GMM, MFCC-based HMM and phase-based GMM, we obtained the correct rates of 99.4%, 98.9%, 98.9% for normal, fast and slow speaking modes, respectively. These results show the best performance in comparison with the other researcher’s results for the same database [1, 6, 15, 16, 17].

To demonstrate the robustness of phase information for speaker recognition, the new phase information \( \{\cos \theta, \sin \theta\} \) was also used in speaker verification. The experiments of the combination showed the equal error rate of 0.18% for normal, 0.37% for fast and 0.71% for slow speaking modes, respectively. These results are the error reduction rate of about 50% in comparison with [19].

6. REFERENCES