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

Longbiao Wang[1]*, Shinji Ohtsuka*[2]*, Seiichi Nakagawa*[2]

[1]Department of Systems Engineering, Shizuoka University, Japan
[2]Department 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 modified feature parameter (that is, coordidates on an unit circle) obtained from the original phase information, and evaluated it by using speech database consisting of normal, fast and slow speaking modes. The speaker identification experiments were performed using NTT database which consists of sentences uttered by 35 Japanese speakers (22 males and 13 females) on five sessions over ten months. Each speaker uttered only 5 training utterances at a normal speaking mode (about 20 seconds in total). The proposed new phase information was more robust than the original phase information for all speaking modes. By integrating the new phase information with the MFCC, 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 Transfrom 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 commerical 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

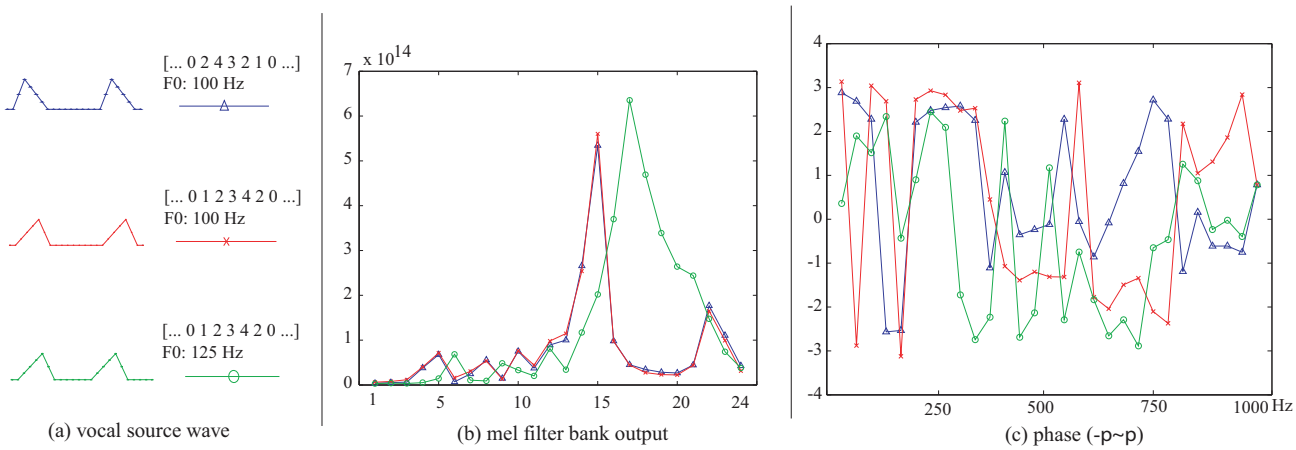
$$\begin{aligned} S(\omega,t) &= X(\omega,t)+jY(\omega,t) \\ &= \sqrt{X^2(\omega,t)+Y^2(\omega,t)}\times e^{j\theta(\omega,t)}, \end{aligned} \quad (1)$$

However, the phase $\theta(\omega,t)$ changes dependending on the clipping position of the input speech even with a same frequency ω. To overcome this problem, the phase of a certain basis frequency ω is kept constant, and the phase of other frequency is estimated relatively. For example, setting the basis frequency ω to $\pi/4$, we have

$$S^{'}(\omega,t)=\sqrt{X^2(\omega,t)+Y^2(\omega,t)}\times e^{j\theta(\omega,t)}\times e^{j\theta(\frac{\pi}{4}-\theta(\omega,t))}, \quad (2)$$

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

$$\begin{aligned} S^{'}(\omega',t) &= \sqrt{X^2(\omega',t)+Y^2(\omega',t)}\times e^{j\theta(\omega',t)}\times e^{j\frac{\omega'}{\omega}(\frac{\pi}{4}-\theta(\omega,t))} \\ &= \tilde{X}(\omega',t)+j\tilde{Y}(\omega',t), \end{aligned} \quad (3)$$



(a) vocal source wave (b) mel filter bank output (c) phase (-p~p)

Fig. 1. Source wave, mel filter bank output and phase of synthesized speech

with this, the phase can be normalized. Then, the real and imaginary part of Equation (3) becomes

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

$$\tilde{Y}(\omega',t) = \sqrt{X^2(\omega',t)+Y^2(\omega',t)} \times \sin\{\theta(\omega',t) + \frac{\omega'}{\omega}(\frac{\pi}{4} - \theta(\omega,t))\}. \quad (5)$$

In the experiments of this paper, the basis frequency ω 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 $\theta_2 = -\pi + \theta_1$, 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 coordidates on a unit circle, that is,

$$\theta \rightarrow \{\cos\theta, \sin\theta\} \quad (6)$$

2.2. Examples

We generated speech wave by using speech synthesis simulator "VT-Calcs" [13], which can control the vocal soure 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/and 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_2^n given by

$$L_2^n = (1-\alpha)L_{MODEL1}^n + \alpha L_{MODEL2}^n \quad (7)$$

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

$$L_3^n = (1-\beta)L_{MODEL1}^n + \beta\{(1-\alpha)L_{MODEL2}^n + \alpha L_{MODEL3}^n\} \quad (8)$$

where L_{MODEL}^n is a likelihood produced by the *n-th* speaker *MODEL*, $n = 1, 2, \cdots, N$, where N is the number of speakers registered. α and β 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 other four sessions were uttered at normal, fast and

Table 1. Speaker identification result (%)

speed	normal	fast	slow	Ave.
MFCC_GMM	97.7	95.1	94.9	95.9
$\{\theta\}$	52.6	51.6	51.7	52.0
$\{\cos\theta, \sin\theta\}$	73.4	72.0	70.4	71.9
MFCC_GMM+$\{\theta\}$	99.0	97.0	96.6	97.5
MFCC_GMM+$\{\cos\theta, \sin\theta\}$	99.3	98.0	98.1	98.5
MFCC_GMM+MFCC_HMM	99.3	97.9	96.6	97.9
MFCC_GMM+MFCC_HMM+$\{\theta\}$	99.4	98.6	97.4	98.5
MFCC_GMM+MFCC_HMM + $\{\cos\theta, \sin\theta\}$	99.4	98.9	98.9	99.1

Table 2. Equal error rate of speaker verification (%)

speed	normal	fast	slow	Ave.
MFCC_GMM	0.58	1.28	1.38	1.08
MFCC_HMM	0.70	1.00	1.43	1.05
$\{\cos\theta, \sin\theta\}$	5.22	6.15	5.69	5.69
MFCC_GMM+MFCC_HMM	0.35	0.91	1.16	0.81
MFCC_GMM+$\{\cos\theta, \sin\theta\}$	0.38	0.75	0.77	0.63
MFCC_HMM+$\{\cos\theta, \sin\theta\}$	0.28	0.40	0.83	0.50
MFCC_GMM+MFCC_HMM + $\{\cos\theta, \sin\theta\}$	0.18	0.37	0.71	0.42

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 [1] having diagonal covariance matrices were used as speaker 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]. That is to say, only the original phase $\{\theta\}$ and the combination of the phase $\{\theta\}$ based GMM and MFCC-based GMM were used to perform the speaker identification on speech data with normal speaking mode [2]. The method *phase* $\{\theta\}$ means that the phase value obtained by Equation (3) was used as speaker identification feature. The method phase $\{\cos\theta, \sin\theta\}$ means that the phase values are transformed to coordinates by Equation (6). So the number of parameters becomes two times in comparison with phase $\{\theta\}$. The new phase $\{\cos\theta, \sin\theta\}$ significantly outperformed the original phase $\{\theta\}$. Although *phase* based method worked worse than MFCC based method, it had some ability of speaker identification. So it might be useful to use phase information to identify the speaker.

The combination method achieved a relative error reduction rate of 52.2% from MFCC based method in the case of new phase information for a normal speed, 55.6% reduction for a slow speed, and 59.2% reduction for a fast speed, respectively. The proposed new phase information $\{\cos\theta, \sin\theta\}$ is more robust than the original phase information θ for all speaking modes. We can guess that the speaker dependent phase feature presents the characteristics of vocal source. The combination of three methods: MFCC-based HMM (116 syllable-based left-to-right HMM having 4 states adapted from speaker independent models, 16 mixtures/state), MFCC-based GMM and phase-based GMM was also performed. The combination of the former two methods was very effective [1]. When the combination of MFCC-based GMM and MFCC-based MM was integrated with the new phase information, a relative error reduction of 57.1% (from 97.9% to 99.1%) was achieved. In other words, by using the new phase information, a 3.2% improvement(78.0% relative error reduction rate) over MFCC-based GMM was achieved.

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 registed speakers was not very large. To evidence the robustness of phase information for speaker recognition, the new phase information $\{\cos\theta, \sin\theta\}$ [3] is used to perform speaker verification[4] in this section.

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-basd GMM and MFCC-based HMM is same as that of speaker identification. The speech analysis conditions are also same as that of speaker identification.

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 $\{\theta\}$ was significantly worse than that of the new phase information $\{\cos\theta, \sin\theta\}$, 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].

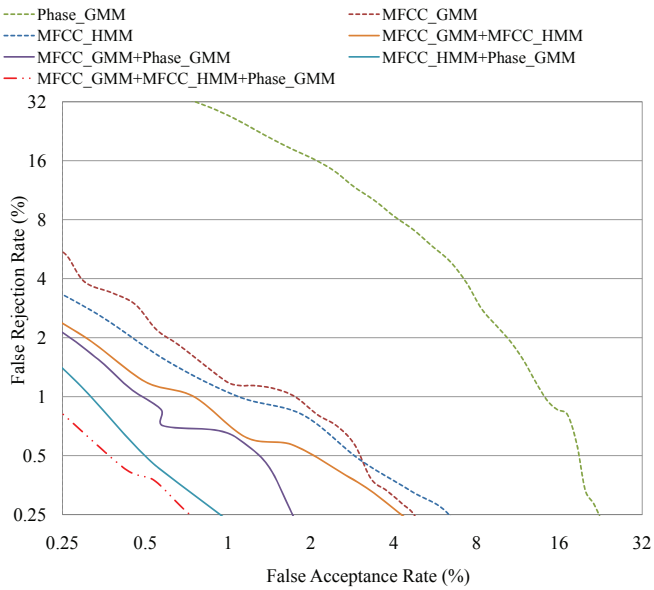


Fig. 2. Comparison of speaker verification performance with different feature parameters by DET curves

$\{\cos\theta, \sin\theta\}$ 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 θ 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, respectivey. These results show the best performance in comparison with the other rescarcher's results for the same database [1, 6, 15, 16, 17].

To demenstrate 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

[1] S.Nakagawa,W.Zhang,M.Takahashi,"Text-independent speaker recognition by combining speaker specific GMM with speaker adapted syllable-based HMM", in Proceeding of ICASSP ,Vol I,pp81-84,2004

[2] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance", in proceedings of ICASSP', Vol. 1, pp. 133–136, 2001.

[3] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception", in proceedings of Eurospeech-2003, pp. 2117–2120, 2003.

[4] G.Shi et.al, "On the importance of phase in human speech recognition", IEEE Trans. Audio, Speech and Language Processing, Vol.14, No.5, pp1867-1874(2006)

[5] P.Aarabi et.al : "Phase-based speech processing", World Scientific (2005)

[6] K.P. Markov and S. Nakagawa, "Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition", Jour. ASJ (E), Vol.20, No.4, pp.281-291 (1999)

[7] M.D. Plumpe, T.F. Quatieri, D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification", IEEE Trans. Speech and Audio Processing, Vol.7, No.5, pp. 569-586 (1999)

[8] K.S.R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker verification", IEEE Signal Processing Letters, Vol.13, No.1, pp.52-55 (2006)

[9] N. Zheng, T. Lee and P.C. Ching, "Integration of complementary acoustic features for speaker recognition", IEEE Signal Processing Letters, Vol.14, No.3, pp.181-184 (2007)

[10] S. Nakagawa, K. Asakawa and L. Wang, "Speaker recognition by combining MFCC and phase information", Proc. Interspeech, pp. 2005-2008, 2007.

[11] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, Vol. 17, No. 1-2, pp. 91–108, 1995.

[12] F. Bimbot et al., "A tutorial on text-independent speaker verification", EURASIP Journal on Applied Signal Processing 2004:4, pp. 430–451, 2004.

[13] http://speechlab.bu.edu/VTCalcs.php

[14] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification", in Proc. International Conf. on Spoken Language Processing (ICSLP '92), vol. 1, pp. 599–602, 1992.

[15] Matusi, T., Furui, S., "Concatenated phoneme models for text-variable speaker recognition", in proceedings of ICASSP'93, Vol.II, pp. 391–394, 1993.

[16] H. Yamamoto, Y. Nankaku, C. Miyajima, K.Tokuda, T. Kitamura, "Parameter sharing in mixture of factor analyzes for speaker identification", IEICE Trans., Vol.E-88D, No.3, pp. 418–424, 2005.

[17] S. Nakagawa, W. Zhang, M. Takahashi, "Text-independnt/text-prompted speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM", IEICE Trans., Vol.E89-D, No.3, pp. 1058–1064, 2006.

[18] A. E. Rosenberg, "Automatic speaker verification: a review", Proc. IEEE, Vol. 64, pp. 475–487, 1976.

[19] K. P. Markov, S. Nakagawa, "Text-independent speaker recognition using non-linear frame likelihood transformation", Speech Communication, Vol. 24, No. 3, pp. 193–209, 1998.