Speaker Recognition by Combining MFCC and Phase Information

Seiichi Nakagawa, Kouhei Asakawa, Longbiao Wang

Department of Information and Computer Sciences, Toyohashi University of Technology, Japan

{nakagawa, asakawa, wang}@slp.ics.tut.ac.jp

Abstract

In conventional speaker recognition method based on MFCC, the phase information has been ignored. In this paper, we proposed a method that integrates the phase information on a speaker recognition method. The speaker identification experiments were performed using NTT database which consists of sentences uttered at normal speed mode by 35 Japanese speakers (22 males and 13 females) on five sessions over ten months. Each speaker uttered only 5 training utterances (about 20 seconds in total). Using the phase information, the speaker recognition error rate was reduced by about 44%.

Index Terms: speaker identification, MFCC, phase information, GMM, 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 of result has been obtained using continuous density HMM (CHMM) for modeling the speaker characteristics [1]. 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 [2]. 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 [3]. In this paper, we use GMM to model speaker characteristics.

Several studies have indicated a large effort to directly model and incorporate the phase into the recognition process [4, 5]. The importance of phase in human speech recognition has been reported in [6, 7]. 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 utilize 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 [8, 9, 10, 11]. In this paper, the phase information is individually used to identify the speaker, and it is also integrated with the MFCC feature.

2. Phase Information Analysis

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

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

However, the phase $\theta(\omega, t)$ changes dependending on the cutting position 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

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

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

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

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

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

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

In the experiments of this paper, the basis frequency ω is set to $2\pi \times 1000\, Hz$. To reduce the number of features parameters, we used only phase information in a sub-band frequency range.

3. Speaker Modeling by Gaussian Mixture Model (GMM)

A GMM is a weighted sum of M component densities and is given by

$$P(x|\lambda) = \sum_{i=1}^{M} c_i b_i(x), \tag{6}$$

where x is a d-dimensional random vector, $b_i(x)$, $i = 1, \cdots, M$, is the component density and c_i, $i = 1, \cdots, M$, is the mixture weight. Each component density is a d-variate Gaussian function given by

$$b_i(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\}, \tag{7}$$

with a mean vector μ_i and a covariance matrix Σ_i. The mixture weights satisfy the following constraint

$$\sum_{i=1}^{M} c_i = 1. \tag{8}$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{c_i, \mu_i, \Sigma_i\}, \quad i = 1, \ldots, M. \tag{9}$$

In our speaker recognition system, each speaker is represented by such a GMM and is referred to by his/her model λ.

For a sequence of T test vectors $X = x_1, x_2, \cdots, x_T$, the standard approach is to calculate the GMM likelihood as

$$L(X|\lambda) = \log p(X|\lambda) = \sum_{t=1}^{T} \log p(x_t|\lambda). \tag{10}$$

The speaker specific GMM parameters are estimated by the E-M algorithm using only the speaker specific training data (the HTK toolkit [12]).

4. Combination Method

In this paper, the GMM based on MFCC is combined with the GMM based on phase information or/and GMM based on logarithmic power spectrum. The *s-th* speaker-specific GMM based on MFCC produces log-likelihood L^s_{MFCC}, $s = 1, 2, \cdots, N$. The *s-th* speaker-specific GMM based on phase information or logarithmic power spectrum also produces log-likelihood L^s_{phase} or $L^s_{spectrum}$ by using phase information or logarithmic power spectrum, respectively. When a combination of two GMMs is used to identify the speaker, the likelihood of GMM based on MFCC is linearly coupled with the GMM based on phase information or the GMM based on logarithmic power spectrum, where they are transformed to form the new score L^s given by

$$L^s = (1-\alpha)L^s_{MFCC} + \alpha L^s_{phase}, \tag{11}$$

or

$$L^s = (1-\alpha)L^s_{MFCC} + \alpha L^s_{spectrum}, \tag{12}$$

where α denotes a weighting coefficient.

When three GMMs are used, the likelihood of GMM based on MFCC is combined with those of GMM based on phase information and GMM based on logarithmic power spectrum as:

$$L^s = \beta \times \{(1-\alpha) \times GMM_{MFCC} + \alpha \times GMM_{phase}\} + (1-\beta) \times GMM_{spectrum}. \tag{13}$$

Table 1: The speaker identification result by individual method (# parameters = 12)

feature parameter	identification rate (%)	
	32 mixtures	64 mixtures
MFCC	95.7	95.9
phase (60-700Hz)	41.0	54.7
phase (300-1000Hz)	33.2	38.1
phase (600-1300Hz)	17.0	23.1
spectrum (60-700Hz)	76.4	77.8
spectrum (300-1000Hz)	65.1	64.4
spectrum (600-1300Hz)	60.8	61.4

5. Experiments

5.1. 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 [13]. 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 $\times$ 4 other sentences from the other four sessions were used as test data. The average duration of the sentences is about 4 seconds. The input speech was sampled at 16 kHz. The spectrum with 128 components consisting of magnitude and phase was calculated by DFT for every 256 samples. 12 MFCCs were calculated at every 5 ms with a window of 12.5 ms.

5.2. Experimental Results

We evaluated the speaker recognition experiment using phase information. GMMs with 32 mixtures or 64 mix-

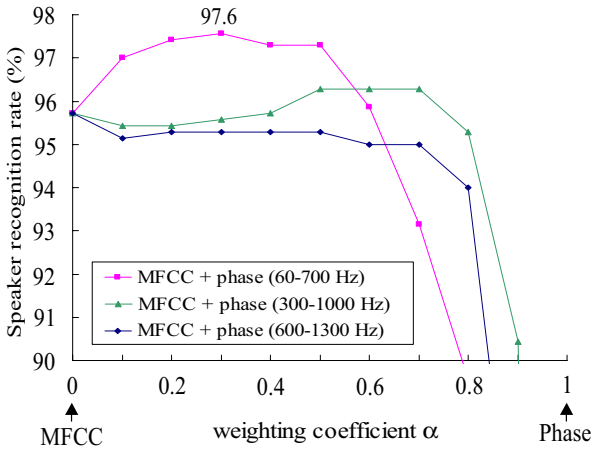

Figure 1: Speaker identification result by combining MFCC with phase information (GMM with 32 mixtures)

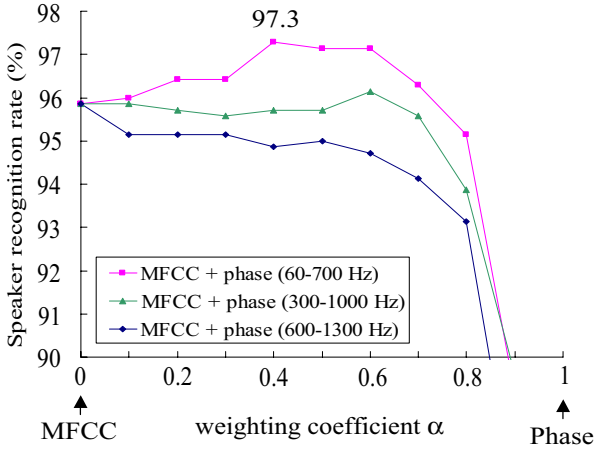

Figure 2: Speaker identification result by combining MFCC with phase information (GMM with 64 mixtures)

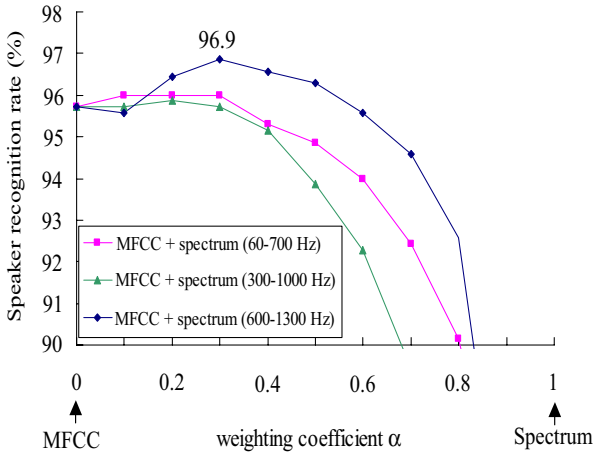

Figure 3: Speaker identification result by combining MFCC with power spectrum (GMM with 32 mixtures)

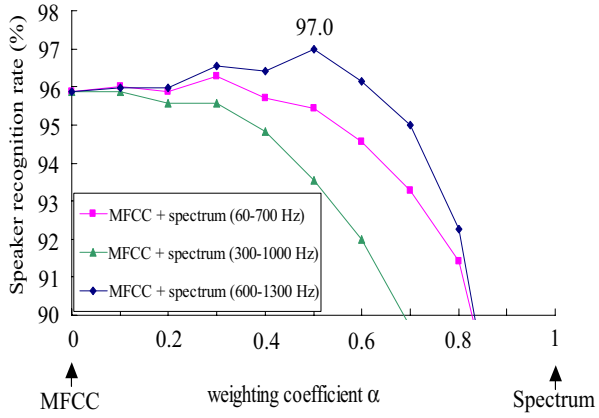

Figure 4: Speaker identification result by combining MFCC with power spectrum (GMM with 64 mixtures)

tures having diagonal covariance matrices were used as speaker models.

The speaker identification result by individual method is shown in Table 1. The method *phase* means that the phase value obtained by Equation (3) was used as speaker recognition feature. "*Phase* (60 Hz - 700 Hz)" corresponds to the 12 feature parameters, that is, from the $1st$ component to $12th$ component of the spectrum. In comparison, we also used "*power spectrum*" in a sub-band frequency range obtained from Equation (3), that is, $\sqrt{X^2(\omega', t) + Y^2(\omega', t)}$. Although *phase* based method worked worse than MFCC, it had some ability of speaker identification. So it might be useful to use phase information to identify the speaker.

The speaker identification results by combining MFCC with phase information are shown in Figs. 1 and 2. The combination method achieved a relative error reduction rate of 44.2% from MFCC based method in the case of phase information from 60 Hz to 700 Hz (GMM with 32 mixtures). Figs. 3 and 4 show the results of the combination of MFCC and power spectrum in a sub-band frequency range. This combination also improved the recognition performance, but the effect is smaller than that of phase information. Finally, we combined the MFCC, phase information and power spectrum at the same time. Figs. 5 and 6 illustrate the results. The combination method further improved the recognition performance. We could reduce the error rate of 53.5% in comparison with the baseline of MFCC (GMM with 32 mixtures).

The other combination method which constructs a GMM using the MFCC feature and phase information was also conducted in this paper. 12-dimension MFCC and 12-dimension phase information were concatenated to a 24-dimension feature. GMMs based on the 24-dimension feature instead of those based on the 12-dimension MFCC were trained and tested. 96.7% speaker idenfication rate was achieved by the GMMs with 32

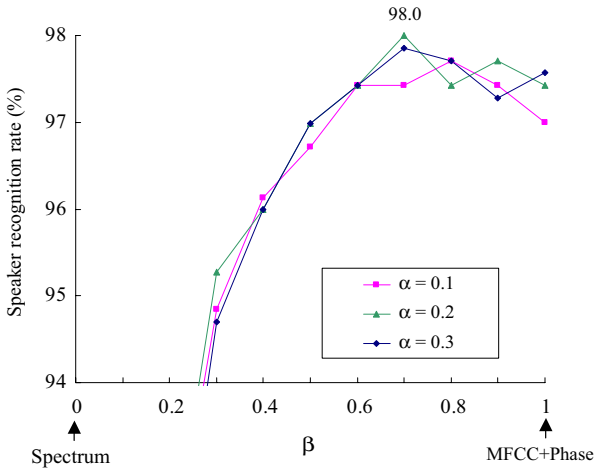

Figure 5: Speaker identification result by combining MFCC with phase information and power spectrum (GMM with 32 mixtures)

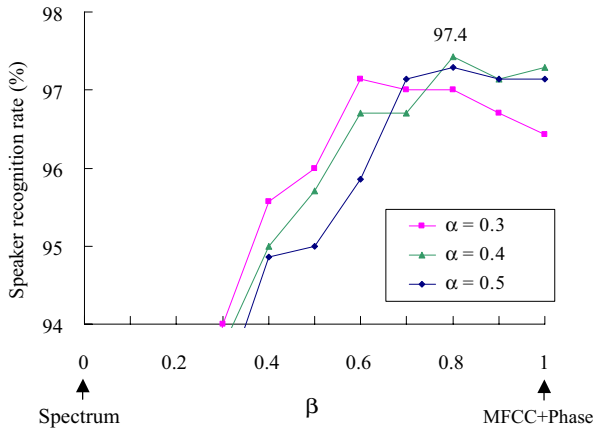

Figure 6: Speaker identification result by combining MFCC with phase information and power spectrum (GMM with 64 mixtures)

mixtures. While the GMMs based on the concatenated parameters did not outperform the combination method described in Section 4 which combined the likelihoods of two independent GMMs, it outperformed the GMMs based on the 12-dimension MFCC (95.7%).

6. Conclusion

We proposed a text-independent speaker recognition method by combining MFCC and phase information. The speaker identification experiments were conducted on NTT database which consists of sentences data uttered at normal speed mode by 35 Japanese speaker. Combining the MFCC and phase information, we obtained the error reduction rate of 44.2% than MFCC. After adding power spectra in 600-1300 Hz, we obtained the error reduction rate of 53.5%.

7. References

[1] Savic, M., Gupta, S., "Variable parameter speaker verification system based on Hidden Markov Modeling, in proceedings of ICASSP'90, pp. 281–284, 1990.

[2] Tseng, B., Soong, F., Rosenberg, A., "Continuous probabilistic acoustic map for speaker recognition", in proceedings of ICASSP'92, vol.II, pp. 161–164, 1992.

[3] D. A. Reynolds, R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans. Speech and Audio Processing, Vol. 3, No. 1, pp. 72–83, 1995.

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

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

[6] 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)

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

[8] 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)

[9] 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)

[10] 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)

[11] 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)

[12] Young, S., Kershow,D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., "The HTK Book", 2000.

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