Speaker Identification Using Pseudo Pitch Synchronized Phase Information in Voiced Sound

Kohta Shimada*, Kazumasa Yamamoto* and Seiichi Nakagawa*
* Department of Computer Science and Engineering, Toyohashi University of Technology, Japan
E-mail:{shimada, kyama, nakagawa} @slp.cs.tut.ac.jp

***Abstract*—In conventional speaker identification methods based on mel-frequency cepstral coefficients (MFCCs), phase information is ignored. Our recent studies have shown that phase information contains speaker dependent characteristics. We propose a new extraction method to extract pitch synchronous phase information from the voiced section only. Speaker identification experiments were performed using the NTT clean database and JNAS database. Using the new phase extraction method, we obtained a relative reduction in the speaker error rate of approximately 27% and 46%, respectively, for the two databases. We also obtained a relative error reduction of approximately 52% and 42%, respectively, when combining phase information with the MFCC-based method.**

I. Introduction

In conventional speaker identification methods based on mel-frequency cepstral coefficients (MFCCs), only the magnitude of the Fourier Transform in time-domain speech frames is used. This means that the phase component is ignored. Of course, MFCCs capture not only speaker-specific vocal tract information, but also vocal source characteristics. Nevertheless, feature parameters extracted from excitation source characteristics are also useful for speaker identification [1], [4], [5], [6], [7], [10]. Almost all of the existing methods are based on Linear Predictive Coding (LPC) analysis. Markov and Nakagawa proposed a Gaussian Mixture Model (GMM) based text-independent speaker identification system that integrates pitch and the LPC residual with the LPC-derived cepstral coefficients [4]. Their experimental results show that using pitch information is the most effective when the correlation between pitch and the cepstral coefficients is taken into consideration. An automatic technique for estimating and modeling the glottal flow derivative source waveform of speech and applying the model parameters to speaker identification was proposed in [5]. The complementary nature of speaker-specific information in the residual phase compared with the information in conventional MFCCs was demonstrated in [6]. The residual phase was derived from speech signals by linear prediction analysis. Zheng et al. proposed a speaker verification system using complementary acoustic features derived from vocal source excitation and the vocal-tract system [7]. A new feature set, called the wavelet octave coefficients of residues (WOCOR), was proposed to capture the spectro-temporal source excitation characteristics embedded in the linear predictive residual signal [7]. Recently, many speaker recognition studies using group delay based phase information have been proposed [8], [9]. Wang et al. proposed phase-related features for speaker recognition [11]. This type of phase information considers all frequency ranges. We think that phase information is valid for speaker identification, since it captures the features of the source wave.

Previously, we proposed a speaker identification system using a combination of MFCCs and phase information [1], [2], directly extracted from the limited bandwidth of the Fourier transform of the speech wave. We also showed that the phase information is effective for speaker identification in clean and noisy environments [1], [2], [3]. However, problems occurred in extracting the phase information because of the influence of the windowing position. Therefore, we propose a new method to extract pitch synchronous phase information in voiced sound only. Using the new extraction method, the speaker identification rate improved by approximately 27% and 46% for the NTT and JNAS databases, respectively.

The rest of this paper is organized as follows. Section 2 presents the phase information extraction method, while Section 3 discusses combining the phase and MFCC methods. The experimental setup and results are reported in Section 4, and Section 5 presents our conclusions.

II. Phase information extraction

A. Formulas [1], [3]

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 changes, depending on the clipping position of the input speech even at the same frequency ω. To overcome this problem, the phase of a certain basis frequency ω is kept constant, and the phases of other frequencies are estimated relative to this. For example, by setting the basis frequency ω to $\pi/4$, we obtain

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

whereas for the other frequency $\omega' = 2\pi f'$, the spectrum becomes

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

In this way, the phase can be normalized. Then, the real and imaginary parts of (3) become

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

In the experiments described in this paper, the basis frequency ω is set to $2\pi \times 1000Hz$. In a previous study, to reduce the number of feature parameters, we used phase information in a sub-band frequency range only. However, a problem arose with this method when comparing two phase values. For example, for two values $\pi - \theta_1$ and $\theta_2 = -\pi + \theta_1$, the difference is $2\pi - 2\theta_1$. If $\theta_1 \approx 0$, then the difference $\approx 2\pi$, despite the two phases being very similar to each other. Therefore, we modified the phase into coordinates on a unit circle [3], that is,

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

B. Improvement of phase information extraction

Using the relative phase extraction method that normalizes the phase variation by cutting positions, we can reduce the phase variation. However, the normalization of phase variation is still inadequate. For example, for a 1000 Hz periodic wave (16 samples per cycle for a 16 kHz sampling frequency), if one sample point shifts in the cutting position, the phase shifts only $\frac{2\pi}{16}$, while for a 500 Hz periodic wave, the phase shifts only $\frac{2\pi}{32}$ with this single sample cutting shift. On the other hand, if the 17 sample points shift, their phases will shift by $\frac{17\cdot 2\pi}{16}(mod 2\pi) = \frac{2\pi}{16}$ and $\frac{34\pi}{32}$, respectively, for the two periodic waves. Therefore, the values of the relative phase information for different cutting positions are very different from those of the original cutting position. We have addressed such variations using a statistical distribution model of GMM [1], [2], [3].

If we could split the utterance by each pitch cycle, changes in the phase information would be further obviated. Thus, we propose a new extraction method that synchronizes the splitting section with a pseudo pitch cycle.

With respect to how to unite the cutting sections in the time domain, the proposed method looks for the maximum amplitude at the center around the conventional target splitting section of an utterance waveform, and the peak of the utterance waveform in this range is adopted as the center of the next window. Fig. 1 outlines how to synchronize the splitting section.

C. Using only the voiced speech section

We think that the proposed phase information is useful for speaker identification because it captures the features of the source waveform. Thus, we extracted phase information from the voiced speech section only. If the energy of the lower spectral components (0 [Hz] $\sim$ 2000 [Hz]) in a given frame is larger than that of the upper spectral components (2000 [Hz] $\sim$ 8000 [Hz]) and the frame's power is larger than the average of all the frame's powers, we judge it to be a voice frame.

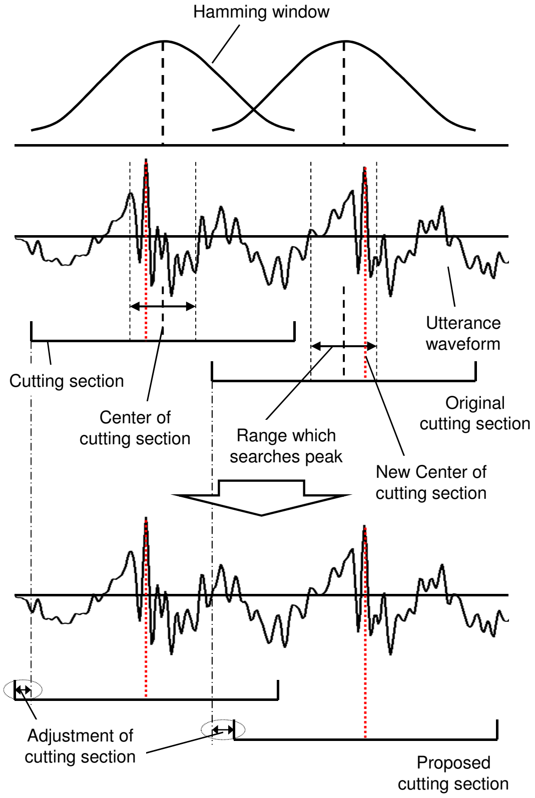

Hamming window
Utterance waveform
Cutting section
Center of cutting section
Range which searches peak
Original cutting section
New Center of cutting section
Adjustment of cutting section
Proposed cutting section

Fig. 1. *How to synchronize the splitting section.*

III. Combination method and decision method

In this paper, the GMM based on MFCCs is combined with the GMM based on phase information. When a combination of the two methods is used to identify the speaker, the likelihood of the MFCC-based GMM is linearly coupled with that of the GMM based on phase information to produce a new score L^n_{conb} given by

$$L^n_{conb} = (1-\alpha)L^n_{MFCC} + \alpha L^n_{phase}, \quad n = 1, 2, \cdots, N, \quad (7)$$

where L^n_{MFCC} and L^n_{phase} are the likelihoods produced by the n-th MFCC-based speaker model and phase information based speaker model, respectively. N is the number of speakers registered and α denotes the weighting coefficients, which are determined empirically. The speaker (or speaker model) with maximum likelihood is judged to be the target speaker.

IV. Experiments

A. Databases and speech analysis

We used the NTT (Nippon Telegraph and Telephone) and JNAS (Japanese Newspaper Article sentence) databases in the experiments. The NTT clean database consists of recordings of 35 speakers (22 males and 13 females), collected in five sessions over 10 months (1990.8, 1990.9, 1990.12, 1991.3, and 1991.6) in a sound-proof room [1], [3], [4]. To train the models, the same five sentences were used for all speakers in one session (1990.8). These sentences were uttered at a normal speaking rate. Five different sentences at each of the other four sessions were uttered at normal, fast, and slow speaking rates and used as test data. In total, the test corpus consisted of 2100 trials (5×4×3×35) for speaker identification. The average duration of the sentences was approximately four seconds.

The JNAS corpus consists of the recordings of 270 speakers (135 males and 135 females). To train the models, 10 sentences were used for all speakers. About ninety other sentences were used as test data. In total, the test corpus consisted of about 24,000 (90×270) trials for speaker identification. The average duration of the sentences was approximately three seconds.

The input speech was sampled at 16 kHz. A total of 25 dimensions (12 MFCCs, 12 ΔMFCCs and Δpower) in the both of JNAS database and NTT database were calculated every 10 ms with a window of 25 ms. We also conducted the experiment by using feature parameters of 12 dimensions (12 MFCCs), because the amount of training data is smaller for its database. Unless otherwise noted, the experimental results described below correspond to the case of 25 dimensions.

The spectrum with 128 components consisting of magnitude and phase was calculated by DFT for every 256 samples. For phase information, we used the first 12 phase components (24 feature parameters in total), that is, from the first to the 12th component of the phase spectrum (frequency range: 60 Hz - 700 Hz), which achieved the best identification performance in all the other sub-band frequency ranges [1].

A frame is judged to be a voiced sound frame, if the ratio of the lower spectral components to the upper spectral components is greater than 4 and the ratio of the frame's power to the average of all the frames' powers is greater than 1/5. Under this condition, about 70% of all the frames were judged to be voiced sound frames.

B. Speaker identification results for NTT database

We conducted a speaker identification experiment using phase information in the NTT database. GMMs with 32 mixtures were used as speaker models. The new phase extraction method searches the peak amplitude point in the ranges ± 0.5 ms, ± 2 ms, and ± 5 ms in the center of the next window. The speaker identification results obtained from the individual methods are shown in Tables I, II, and Fig. 2. "0 ms" corresponds to the conventional extraction method in which the recognition rate for normal speed utterances is 74.7%. On the other hand, using the newly proposed extraction technique, the recognition rate improves to 81.6%. Moreover, for the slow speed, the rate using the proposed method (voiced sound, ± 2ms) improved from 72.1% to 80.7%. Similarly, for fast speed, the rate improved from 73.0% to 79.3%. Overall, the average error reduction rate is 27.0% (73.3% to 80.5%). Using the voiced sections only, the error reduction rate is 13.3% (77.5% to 80.5%). By comparing 12 dimensions (MFCC) and 25 dimensions (MFCC), we found that the performance by using 12 dimensions was slightly superior to that of 25 dimensions. It depends on the amount of training data.

This improvement is due to phase information was extracted from only voiced section which vocal cords vibrate periodically. And in the improvement for every speed utterances, these improvements have a relationship with adjusting the cutting position to be roughly synchronized the pitch period.

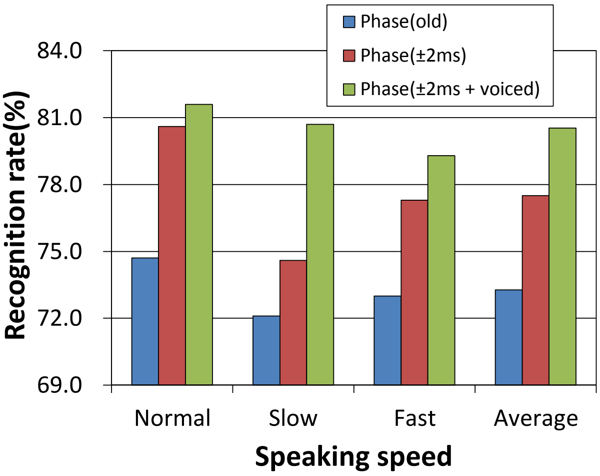


Fig. 2. *Speaker identification results using phase with the proposed method for NTT database.*

TABLE I
Speaker identification results for NTT database (%).

(a) 12 dimensions of feature parameters (MFCC)

	phase / phase & MFCC				MFCC
search range (ms)	0	0.5	± 2	± 5	
normal speed	74.7 / 99.0	73.3 / 99.1	80.6 / 99.4	79.1 / 99.3	98.9
slow speed	72.1 / 97.6	71.7 / 97.9	74.6 / 98.4	76.0 / 98.3	95.7
fast speed	73.0 / 98.1	71.6 / 98.1	77.3 / 98.4	77.6 / 98.7	96.6
average	73.3 / 98.2	72.2 / 98.4	77.5 / 98.7	77.6 / 98.8	97.1

(b) 25 dimensions of feature parameters (MFCC)

	phase / phase & MFCC				MFCC
search range (ms)	0	0.5	± 2	± 5	
normal speed	74.7 / 98.7	73.3 / 99.0	80.6 / 99.0	79.1 / 98.9	97.3
slow speed	72.1 / 97.1	71.7 / 97.6	74.6 / 97.9	76.0 / 97.4	95.4
fast speed	73.0 / 98.1	71.6 / 98.6	77.3 / 98.4	77.6 / 98.6	95.6
average	73.3 / 98.0	72.2 / 98.4	77.5 / 98.4	77.6 / 98.3	96.1

TABLE II
Speaker identification results for NTT database using voiced sound only (%).

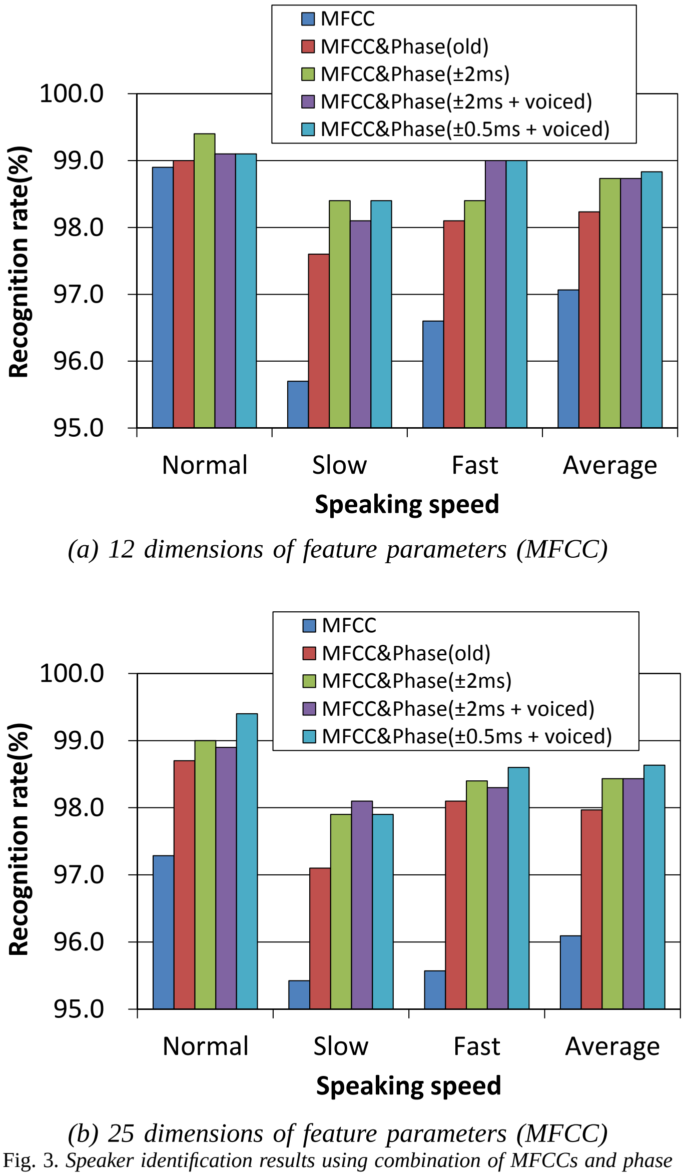
(a) 12 dimensions of feature parameters (MFCC)

	phase / phase & MFCC			
search range (ms)	0	± 0.5	± 2	± 5
normal speed	74.7 / 99.0	74.9 / 99.1	81.6 / 99.1	81.4 / 99.1
slow speed	74.1 / 98.6	74.1 / 98.4	80.7 / 98.1	77.9 / 98.0
fast speed	73.3 / 98.1	75.0 / 98.7	79.3 / 99.0	79.4 / 99.0
average	74.0 / 98.6	74.7 / 98.7	80.5 / 98.7	79.6 / 98.7

(b) 25 dimensions of feature parameters (MFCC)

	phase / phase & MFCC			
search range (ms)	0	± 0.5	± 2	± 5
normal speed	74.7 / 99.0	74.9 / 99.4	81.6 / 98.9	81.4 / 98.7
slow speed	74.1 / 97.9	74.1 / 97.9	80.7 / 98.1	77.9 / 97.1
fast speed	73.3 / 98.4	75.0 / 98.6	79.3 / 98.3	79.4 / 98.7
average	74.0 / 98.4	74.7 / 98.6	80.5 / 98.4	79.6 / 98.2

The speaker identification results obtained from the combination method are shown in Figs. 3-6. The improvement using this method is remarkable. For example, when the MFCC-based method is compared with the combination method, the rate improves from 97.3% to 99.1% (a relative error reduction of 66.7%) for normal speed utterances. By combining phase information with the MFCCs, the average error reduction rate is 64.1% (from 96.1% to 98.6%). This suggests that the proposed phase information extraction method is more effective than the conventional extraction method. Moreover, we clarified that phase information is dependent on the sound source waveform.



(a) 12 dimensions of feature parameters (MFCC)

(b) 25 dimensions of feature parameters (MFCC)

Fig. 3. *Speaker identification results using combination of MFCCs and phase for NTT database.*

C. *Speaker identification results for JNAS database*

We also conducted a speaker identification experiment using phase information in the JNAS database. GMMs with 128 mixtures were used as speaker models. The new phase extraction method searches for the peak amplitude point in the

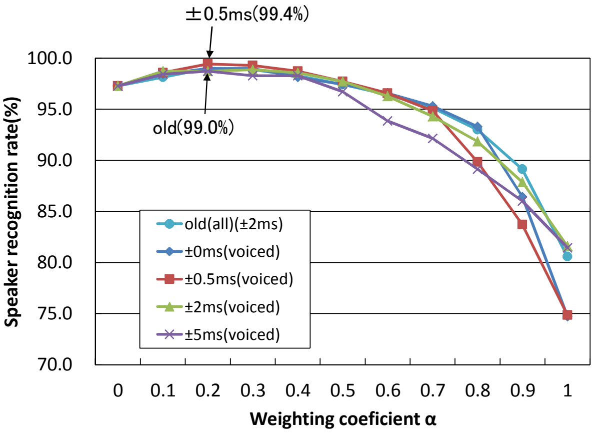


Fig. 4. *Speaker identification results using combination of MFCCs and phase for NTT database (normal speed). -25 dimensions-*

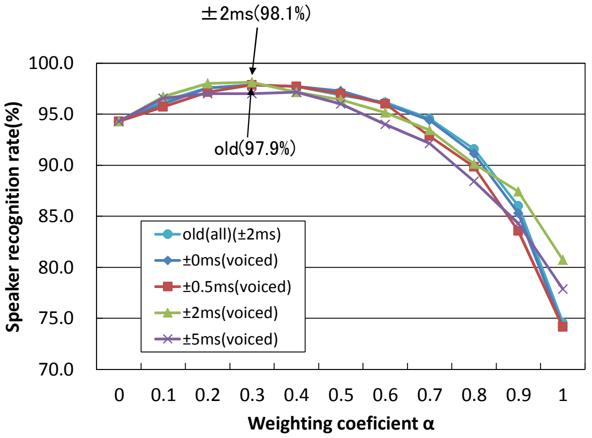


Fig. 5. *Speaker identification results using combination of MFCCs and phase for NTT database (slow speed). -25 dimensions-*

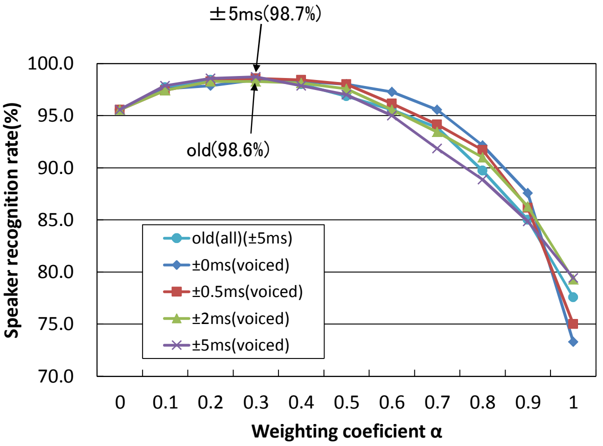


Fig. 6. *Speaker identification results using combination of MFCCs and phase for NTT database (fast speed). -25 dimensions-*

range ±.0.5 ms, ±2 ms, and ±5 ms in the center of the next window. The speaker identification results obtained from the individual methods are shown in Table III and Fig. 7. ”0 ms” corresponds to the conventional extraction method in which the recognition rate is 88.8%. On the other hand, using the newly proposed extraction technique (voiced sound, ±2 ms), the recognition rate improves to 93.9%; that is, an average error reduction rate of 45.5%.

TABLE III
Speaker identification results for JNAS database using voiced sound only (%).

	All sections	Voiced sound only
MFCC	98.8	
search range (ms)	phase / phase & MFCC	
0	88.8 / 99.0	88.3 / 99.1
±0.5	92.8 / 99.2	90.6 / 99.1
±2	92.8 / 99.2	93.9 / 99.3
±5	92.4 / 99.2	93.8 / 99.2

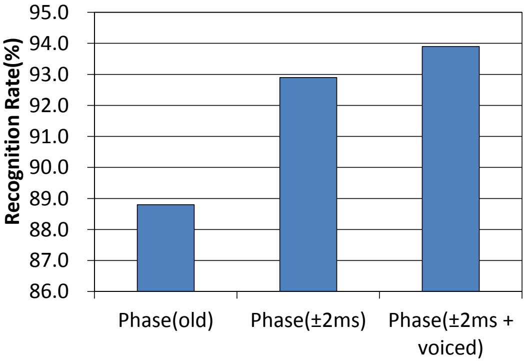


Fig. 7. *Speaker identification results using phase in the JNAS database.*

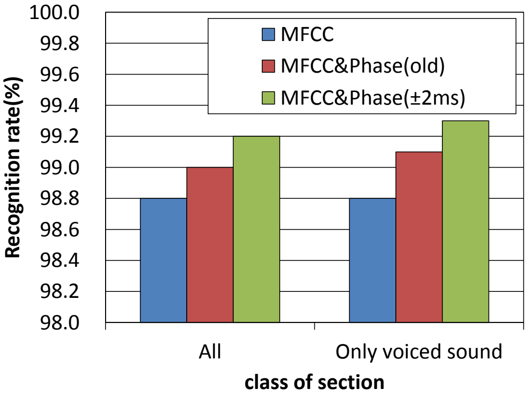


Fig. 8. *Speaker identification results using combination of MFCC and phase in the JNAS database.*

The speaker identification results obtained from the combination method are shown in Figs. 8-10. The improvement using this method is remarkable. For example, when the MFCC-based method is compared with the combination method, the rate improves from 98.8% to 99.1% (a relative error reduction of 25%) for the old phase extraction method, and from 98.8%

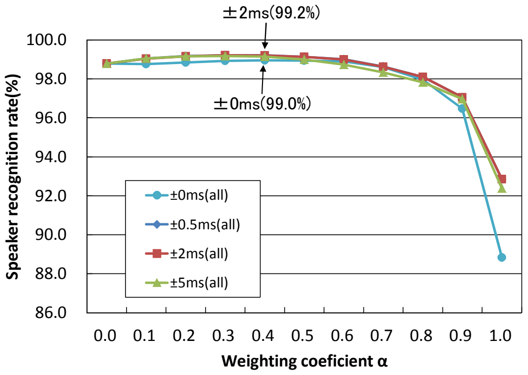


Fig. 9. *Speaker identification results using combination of MFCC and phase in the JNAS database (all sound section).*

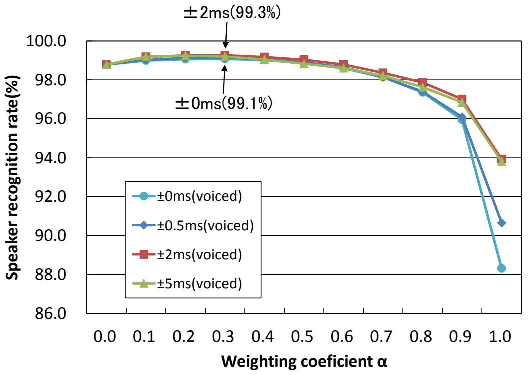


Fig. 10. *Speaker identification results using combination of MFCC and phase in the JNAS database (voiced sound only).*

to 99.3% (a relative error reduction rate of 41.7%) for the new phase extraction method. This result suggests that the proposed phase information extraction method is more effective than the conventional extraction method.

V. Conclusions

In this paper, we proposed a new phase extraction method that extracts pseudo-pitch synchronous phase information from voiced sound sections only. Using the proposed method, the speaker recognition rate for the NTT database improved from 73.3% to 80.5%. Moreover, the recognition rate using the MFCC model improved remarkably when combined with the phase information (from 96.1% to 98.6%). For the JNAS database, the speaker recognition rate improved from 88.8% to 93.9% using only phase information. Moreover, the recognition rate using the MFCC model also improved remarkably when combined with the phase information (from 98.8% to 99.3%). These results confirm that the proposed phase information is most useful for speaker identification.

References

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

[2] L. Wang, S. Ohtsuka, S. Nakagawa, “ High improvement of speaker identification and verification by combining MFCC and phase information ”, Proc. ICASSP, pp.4529-4532, (2009).

[3] L. Wang, K. Minami, K. Yamamoto and S. Nakagawa, “ Speaker identification by combining MFCC and phase information in noisy environments ”, Proc. ICASSP, pp.4502-4505, (2010).

[4] 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).

[5] M.D. Plumpe, T.F. Quatieri and 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).

[6] 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).

[7] 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).

[8] R. Padmanabhan, S. Parthasarathi, H. Murthy, ”Robustness of phase based features for speaker recognition ”, Proc. Interspeech, pp. 2355-2358 (2009).

[9] J. Kua, J. Epps, E. Ambikairajah, E. Choi, ”LS regularization of group delay features for speaker recognition ”, Proc. Interspeech, pp. 2887-2890 (2009).

[10] T. Drugman, T. Dutoit, ”On the potential of glottal signatures for speaker recognition ”, Proc. Interspeech, pp. 2106-2109(2010).

[11] N. Wang, P. C. Ching, T. Lee, ”Exploitation of phase information for speaker recognition ”, Proc. Interspeech, pp. 2126-2129(2010).