Speaker Identification and Verification by Combining MFCC and Phase Information

Seiichi Nakagawa, *Member, IEEE*, Longbiao Wang, *Member, IEEE*, and Shinji Ohtsuka

***Abstract*—In conventional speaker recognition methods based on Mel-frequency cepstral coefficients (MFCCs), phase information has hitherto been ignored. In this paper, we propose a phase information extraction method that normalizes the change variation in the phase according to the frame position of the input speech and combines the phase information with MFCCs in text-independent speaker identification and verification methods. There is a problem with the original phase information extraction method when comparing two phase values. For example, the difference in the two values of $\pi - \tilde{\theta}_1$ and $\tilde{\theta}_2 = -\pi + \tilde{\theta}_1$ is $2\pi - 2\tilde{\theta}_1$. If $\tilde{\theta}_1 \approx 0$, then the difference $\approx 2\pi$, despite the two phases being very similar to one another. To address this problem, we map the phase into coordinates on a unit circle. Speaker identification and verification experiments are performed using the NTT database which consists of sentences uttered by 35 (22 male and 13 female) Japanese speakers with normal, fast and slow speaking modes during five sessions. Although the phase information-based method performs worse than the MFCC-based method, it augments the MFCC and the combination is useful for speaker recognition. The proposed modified phase information is more robust than the original phase information for all speaking modes. By integrating the modified phase information with the MFCCs, the speaker identification rate was improved to 98.8% from 97.4% (MFCC), and equal error rate for speaker verification was reduced to 0.45% from 0.72% (MFCC), respectively. We also conducted the speaker identification and verification experiments on a large-scale Japanese Newspaper Article Sentences (JNAS) database, a similar trend as NTT database was obtained.**

***Index Terms*—Gaussian mixture model (GMM), Mel-frequency cepstral coefficient (MFCC), phase information, speaker identification, speaker verification.**

I. Introduction

OVER the last decade, speaker recognition technology has been used in several commercial products. In this paper, we focus on text-independent speaker recognition. The general field of speaker recognition includes two fundamental tasks: speaker identification and speaker verification [1]–[5]. Speaker identification involves classifying a voice sample as belonging to (that is, having been spoken by) one of a set of N reference speakers (N possible outcomes), whereas speaker verification involves deciding whether or not a voice sample belongs to a specific reference speaker (two possible outcomes—the sample is either accepted as belonging to the reference speaker or rejected as belonging to an impostor).

In conventional speaker recognition methods based on Mel-frequency cepstral coefficients (MFCCs), only the power of the Fourier transform of the time-domain speech frames is used, which means that the phase component is ignored. The importance of phase in human speech recognition has been reported in [6]–[8]. In [6], Liu *et al.* investigated the role of phase information for the human perception of intervocalic plosives. Their results indicate that the short-term amplitude spectra cannot be specified exclusively by plosives. Moreover, the authors concluded that the perception of voicing for plosives relies strongly on phase information. Paliwal and Alsteris also investigated the relative importance of short-time magnitude and phase spectra on speech perception [7]. Human perception experiments were conducted to measure intelligibility of speech tokens synthesized from either the magnitude or phase spectrum. It was shown in [7] that even for shorter windows, the phase spectrum can contribute as much as the magnitude spectrum to speech intelligibility if the shape of the window function is properly selected. In [8], Shi *et al.* analyzed the effects of uncertainty in the phase of speech signals on the word recognition error rate of human listeners. Their results indicate that a small amount of phase error or uncertainty does not affect the recognition rate, but a large amount of phase uncertainty has a significant effect on the human speech recognition rate. Therefore, the phase may also be important in automatic speech/speaker recognition.

Several studies have invested great effort in modeling and incorporating the phase into the speaker recognition process [9]–[11]. The complementary nature of speaker-specific information in the residual phase compared with the information in the conventional MFCCs was demonstrated in [11]. The residual phase was derived from speech signals by linear prediction analysis. Recently, many speaker recognition studies using group delay-based phase information have been proposed [12]–[14]. Group delay is defined as the negative derivative of the phase of the Fourier transform of a signal. In [13], the authors analytically showed why the group delay based phase are robust to noise. The reader can refer to [19] for the explanation of noise robustness of modified group delay. A speaker verification task on the NIST 2003 dataset [35] resulted in better performance for modified group delay features [12] [about 15% equal error rate (EER)] when compared with conventional MFCC features (about 18% EER). In [14], the authors proposed an alternative complementary feature

Manuscript received June 22, 2010; revised February 14, 2011; accepted September 27, 2011. Date of publication October 17, 2011; date of current version February 10, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nestor Becerra Yoma.

S. Nakagawa is with the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi, Aichi, 4418580 Japan (e-mail: nakagawa@slp.ics.tut.ac.jp; nakagawa@slp.cs.tut.ac.jp).

L. Wang is with the Department of Systems Engineering, Shizuoka University, Hamamatsu, Shizuoka 4328561 Japan (e-mail: wang@sys.eng.shizuoka.ac.jp).

S. Ohtsuka is with the Yazaki Syscomplus Co., Ltd., Makinohara 4210501, Japan (e-mail: fuga1023@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2172422

extraction method to reduce the variability of the group delay features derived from the speech spectrum with least squares regularization. The proposed log compressed least square group delay achieved 10.01% EER compared with 7.64% EER for MFCC, and the fusion of group delay and MFCC improved to 7.16% EER for the NIST 2001 SRE database. Evaluations using the NIST 2008 SRE database showed a relative improvement of 18% EER using a group delay-based system combined with a MFCC-based system. Actually, the group delay based phase contains both the power spectrum and phase information [12]–[14], and thus the complementary nature of the power spectrum-based MFCC and group delay phase was not sufficient enough.

We investigate the effect of phase on speaker recognition using both synthesized and human speech. The conclusion reached is that phase information is effective for speaker recognition. However, the phase changes according to the frame position in the input speech. In this paper, we propose a phase information normalization method to address the above problem. There is a problem with this method when comparing two phase values. For example, given two values $\pi - \tilde{\theta}_1$ and $\tilde{\theta}_2 = -\pi + \tilde{\theta}_1$, the difference is $2\pi - 2\tilde{\theta}_1$. If $\tilde{\theta}_1 \approx 0$, then the difference $\approx 2\pi$, despite the two phases being very similar to one another. Therefore, we map the phase into coordinates on a unit circle, that is, $\tilde{\theta} \rightarrow \{\cos\tilde{\theta}, \sin\tilde{\theta}\}$.

With regards speaker recognition, various types of speaker models have been studied over time. The Gaussian mixture model (GMM) has been widely used as a speaker model [2], [4], [16]–[18]. 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 by the capability of Gaussian mixtures to model arbitrary densities. In this paper, an MFCC-based GMM is combined with a phase-based GMM. The proposed phase information is evaluated with respect to both speaker identification and verification.

The remainder of this paper is organized as follows. Section II investigates the importance of phase for speaker recognition and formulates the phase information. Section III briefly describes the combination method for speaker recognition and the speaker identification/verification method. The performances for speaker identification/verification using phase information are evaluated in Section IV. The results of our proposed method are compared with those of related works in Section V. Finally, Section VI summarizes the paper and describes future work.

II. Phase Information Analysis

In this section, we first investigate the effect of phase on speaker recognition, and then formulate the phase information.

A. *Investigating the Effect of Phase*

We investigated the effect of phase on speaker recognition using both synthesized and human speech with various excitation sources and vocal tracts. We generated a speech wave using the speech synthesis simulator "VTCalcs" [20], which can control the voice source wave, pitch (F0) and vocal tract shape. Fig. 1 illustrates the phase and mel filter bank output of the power spectrum for different voice sources and pitch, and a fixed vocal tract shape corresponding to vowel /a/. The solid line with "x" marks, which was the filter bank output of power spectrum obtained from the time-reversed shape[1] for the original shape of voice source [see Fig. 1(a)], was similar to the dash line with triangle marks, which was the output obtained from the original shape where both outputs were obtained from the same F0 and same vocal tract shape. The conventional MFCCs (that is, the DCT of the logarithm of the mel filter bank output) that ignore the phase information cannot capture all the speaker characteristics contained in a voice sources with the same power spectrum, but a different phase. In other words, speaker characteristics in the voice source are not captured completely by the MFCC since the phase information is ignored. As shown in Fig. 1, the phase is greatly influenced by voice source characteristics. Of course, the phase is also influenced by pitch as shown in Fig. 1. In this paper, the distribution of phase for a speaker was modeled by GMMs[2]. Fig. 2 illustrates the mel filter bank output of the power spectrum and phase for the same voice source, but different vocal tracts (Japanese vowels /a/and/i/). The phases are similar and the power spectra are influenced greatly by the vocal tract with the same voice source. To capture the speaker characteristics in the voice source and vocal tract exactly, both power spectrum and phase information of the input speech are required.

We also verified the effect of phase on speaker recognition using human speech. The speech of two Japanese vowels, /a/and/i/, by two Japanese male students was recorded. The mel filter bank output of the power spectrum and the phase information are illustrated in Fig. 3. Compared with the synthesized speech, both the voice source (/a/ uttered by two male students) and the vocal tract (/a/and/i/ uttered by the same person) influence the mel filter bank output of the power spectrum and phase. The reason for this may be that the production of human speech is more complex than that of synthesized speech. Thus, it is difficult to model speaker characteristics using only one set of feature parameters. We expect that the combined usage of MFCC and phase information can help to distinguish the speaker characteristics.

To verify the effectiveness of phase for speaker recognition, inter-speaker and intra-speaker variability of phase was compared. The inter-speaker and intra-speaker variability (Euclidean distance) of phase for five Japanese vowels /a/,/i/,/u/,/e/, and /o/ collected from four speakers is shown in Table I. The results show that phase varies among different vowels for the same speaker. The variability of phase for the same vowel by the same speaker is much smaller than that for the same vowel by different speakers. In other words, a phase information-based GMM is useful for speaker recognition, as the GMM potentially models the speaker characteristics by various vowels (or consonants) corresponding to various Gaussian components and the probability of a test vowel (or consonant) given by the corresponding vowel (consonant) component of the GMM for the target speaker should be much higher than that for the other speakers.

[1] Since "time-reversed shape" is not realistic, we present an artificial example, in which MFCCs cannot discriminate the difference between voice sources with the same power spectrum but different phase.

[2] The GMMs are insensitive to the temporal aspects of speech, they model only the static distribution of features (acoustic observations), but not temporal variations in features from a speaker. Furthermore, one of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrary-shaped densities. Thus, almost all features have been successfully modeled by GMMs with the sufficient number of mixtures.

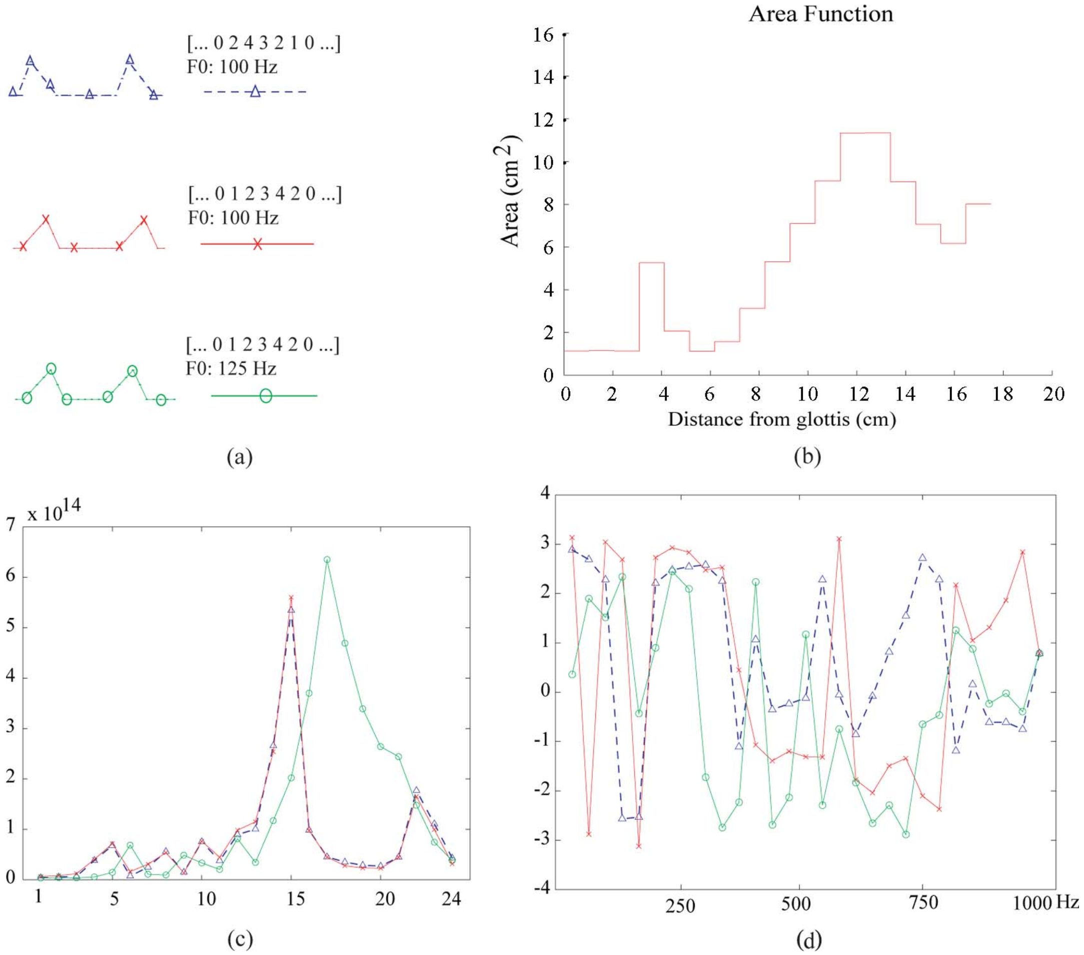


Fig. 1. Source wave, mel filter bank output and phase of synthesized speech for different voice sources and pitch, but the same vocal tract. (a) Vocal source wave. (b) Vocal tract shape of vowel /a/. (c) Mel filter bank output. (d) Phase $(-\pi \sim \pi)$.

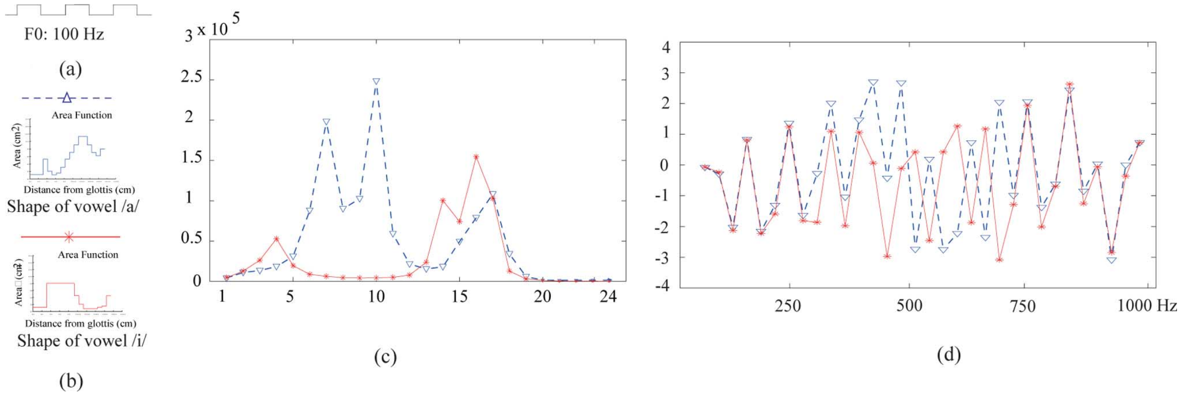


Fig. 2. Voice source, vocal tract, mel filter bank output, and phase of synthesized speech with the same voice source, but different vocal tract. (a) Vocal source. (b) Vocal tract. (c) Mel filter bank output. (d) Phase $(-\pi \sim \pi)$.

B. Formulation of Phase Information

The short-term spectrum $S(\omega, t)$ for the ith frame of a signal is obtained by the DFT of an input speech signal sequence

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

For conventional MFCCs, the power spectrum $X^2(\omega, t) + Y^2(\omega, t)$ is used, but the phase information $\theta(\omega, t)$ is ignored. In this paper, phase $\theta(\omega, t)$ is also extracted as one of the feature parameter set for speaker recognition. The GMMs used in this paper are insensitive to the temporal aspects of speech, and do not capture the dependence of features extracted from each frame. Phase information of the same person with the same voice extracted from different frames may be $\theta(\omega, t)$ and $2\pi + \theta(\omega, t)$. They express different phases value and the different speaker characteristics using phase-based GMMs. In this paper, phase value is constrained to $[-\pi, \pi]$; thus,

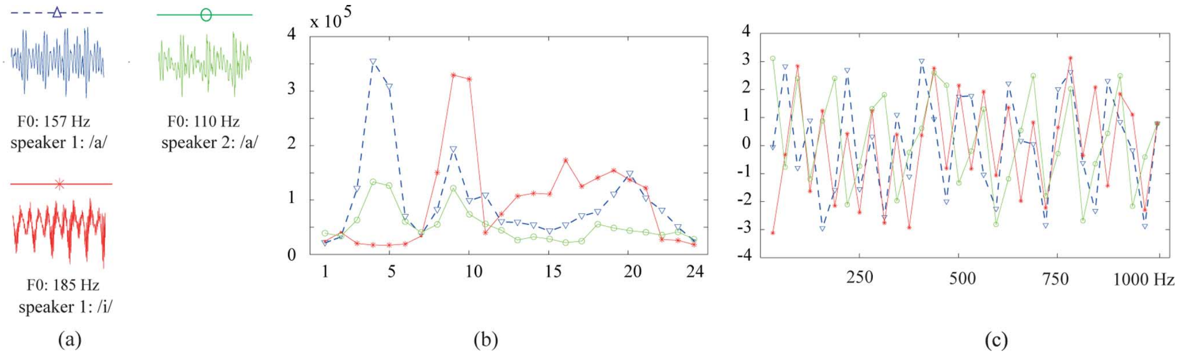


Fig. 3. Source wave, mel filter bank output and phase of human speech (Japanese vowels /a/and/i/). (a) Human speech. (b) Mel filter bank output. (d) Phase $(-\pi \sim \pi)$.

TABLE I
INTER-SPEAKER AND INTRA-SPEAKER VARIABILITIES (EUCLIDEAN DISTANCE) OF PHASE BY FIVE JAPANESE VOWELS COLLECTED FROM FOUR SPEAKERS

	intra-speaker (same speaker)	inter-speaker (different speakers)
intra-vowel (same vowel)	1.29	2.05
inter-vowel (different vowels)	2.68	2.96

$\theta(\omega, t)$ and $2\pi + \theta(\omega, t)$ are converted to the same phase value. Therefore, it is no problem to use GMMs to model the speaker characteristics using phase information.

However, the phase $\theta(\omega, t)$ changes according to the frame position in the input speech. To help the reader to understand the effectiveness of our proposed phase processing, an example of the effect of the frame position on phase for Japanese vowel /a/ is illustrated in Fig. 4. As shown in Fig. 4(b), the unnormalized wrapped phases of two windows become quite a bit different because the phases change according to the frame position. It is obvious that the phase $\theta(\omega, t)$ differs for different frame positions. For speaker recognition using phase information, the phases extracted from two different windows of a same sentence from same people should be as small as possible. Thus, it is necessary to normalize the phase response with respect to the frame position.

A basic method for eliminating the influence of the phase response with respect to frame position is explained as follows. Let $s_1, s_2, \ldots, s_{L_\omega}, \{s_{L_\omega+1} = s_1\}$ be the sampling sequence for a cyclic function, where sample per period (wave length) $L_\omega = f_s/f = 2\pi/\omega f_s$ on radian frequency ω, and f_s is sampling frequency. The phase of $s_1, s_2, \ldots, s_{L_\omega}$ and the phase of $s_2, \ldots, s_{L_\omega}, s_{L_\omega+1}$ are different from each other. The difference of phase on the radian frequency ω is $2\pi/L_\omega$. The phase information is normalized using this attribution in the following. To overcome the influence of the phase response with respect to frame position, phases with the basis radian frequency ω_b for all frames is converted to a constant, and the phase with the other frequency is estimated relative to this. In the experiments discussed in this paper, the basis radian frequency ω_b is set to $2\pi \times 1000$ Hz. Actually, this constant phase value of the basis radian frequency $2\pi \times 1000$ Hz does not affect the speaker recognition result. Without loss of generality, setting the phase with the basis radian frequency $\theta(\omega_b, t)$ to 0, we have

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

The difference between the unnormalized wrapped phase $\theta(\omega_b, t)$ with basis frequency ω_b in (1) and the normalized wrapped phase in (2) is $(-\theta(\omega_b, t))$. With $\omega = 2\pi f \neq 2\pi \times 1000$ Hz, the difference becomes $(\omega/\omega_b)(-\theta(\omega_b, t))$.[3] The spectrum on frequency ω is normalized as

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

Then, the real and imaginary parts of (3) are given by

$$\begin{aligned} \tilde{X}(\omega, t) &= \sqrt{X^2(\omega, t) + Y^2(\omega, t)} \\ &\quad \times \cos\{\theta(\omega, t) + \frac{\omega}{\omega_b}(-\theta(\omega_b, t))\}, \end{aligned} \tag{4}$$

$$\begin{aligned} \tilde{Y}(\omega, t) &= \sqrt{X^2(\omega, t) + Y^2(\omega, t)} \\ &\quad \times \sin\{\theta(\omega, t) + \frac{\omega}{\omega_b}(-\theta(\omega_b, t))\} \end{aligned} \tag{5}$$

and the phase information is normalized as

$$\tilde{\theta}(\omega, t) = \theta(\omega, t) + \frac{\omega}{\omega_b}(-\theta(\omega_b, t)) \tag{6}$$

where it is referred to *the proposed original phase* or *the original normalized phase*.

However, there is a problem with this method when comparing two phase values. For example, with the two values $\pi - \tilde{\theta}_1$ and $\tilde{\theta}_2 = -\pi + \tilde{\theta}_1$, the difference is $2\pi - 2\tilde{\theta}_1$. If $\tilde{\theta}_1 \approx 0$, then the difference $\approx 2\pi$, despite the two phases being very similar to one another. Therefore, for this research, we mapped the phase into coordinates on a unit circle, that is,

$$\tilde{\theta} \rightarrow \{\cos\tilde{\theta}, \sin\tilde{\theta}\}. \tag{7}$$

[3] The different point n is $n = L_{\omega_b} \cdot \theta(\omega_b, t)/2\pi \Leftarrow n \cdot 2\pi/L_{\omega_b} = \theta(\omega_b, t)$. Thus, the difference in phase on radian frequency ω is $L_{\omega_b} \cdot \theta(\omega_b, t)/2\pi \cdot 2\pi/L_\omega = 2\pi/\omega_b f_s/2\pi/\omega f_s \cdot \theta(\omega_b, t) = \omega/\omega_b \cdot \theta(\omega_b, t)$.

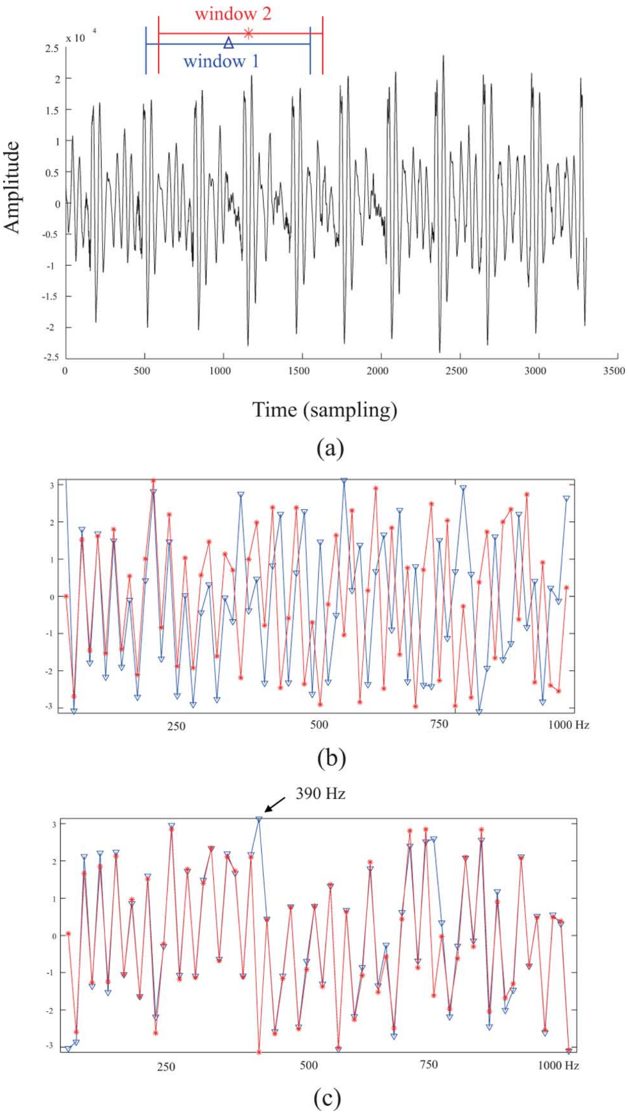


Fig. 4. Example of the effect of frame position on phase for Japanese vowel /a/. (a) Wave form of vowel /a/ and two clipping windows. (b) Unnormalized wrapped phases of two different windows. (c) Normalized wrapped phases of two different windows.

The unnormalized wrapped phase obtained from (1) and the normalized wrapped phase obtained from (6) were compared in Fig. 4(b)–(c). After normalizing the wrapped phase by (6), the phase values shown in Fig. 4(c) become very similar, in other words, the normalized wrapped phase is more adaptable to speaker recognition. The Euclidian distance of unnormalized wrapped phases obtained from (1) and normalized wrapped phases obtained from (6) with 60–1000 Hz sub-band frequency range of two different frame windows were 21.6 and 8.3, respectively. Even the wrapped phase was normalized, it has a problem when comparing two phases which are near π or $-\pi$, the difference was very large despite the two phases being very similar to one another. For example, two unnormalized wrapped phases on 390 Hz were 3.130 and -3.1396, respectively. The difference in two original unnormalized wrapped phases obtained from (6) were 6.2696, nevertheless they should be similar. If the original unnormalized wrapped phase was changed to modified normalized wrapped phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ obtained from (7), the difference became to $\{0, 0.01\}$. Using the modified normalized wrapped phase, the difference in phases extracted from two similar signal segments is very small. In other words, the modified normalized wrapped phases extracted from two different windows of the same utterance from same person represent similar values.

III. Speaker Recognition Method

A. Combination Method

A Gaussian mixture model (GMM) has been widely used as a speaker model [2], [4], [17], [18]. 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 by the capability of Gaussian mixtures to model arbitrary densities. Consideration of the state-of-the-art speaker models is beyond the scope of this paper.

In this paper, the GMM based on MFCCs 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 MFCC-based GMM is linearly coupled with that of the phase information-based GMM to produce a new score L^n_{comb} given by [4], [18]

$$L^n_{\text{comb}} = (1-\alpha)L^n_{\text{MFCC}} + \alpha L^n_{\text{phase}},\ n = 1, 2, \ldots, N \quad (8)$$

where L^n_{MFCC} and L^n_{phase} are the likelihood produced by the nth MFCC-based speaker model and phase information-based speaker model, respectively. N is the number of speakers registered and α denotes weighting coefficients. A speaker with the maximum likelihood is decided as the target speaker.

B. Decision Method

In speaker identification, the speaker with the maximum likelihood is chosen as the target speaker. In speaker verification, the decision whether or not an unlabeled voice sample belongs to a specific reference speaker is made according to the likelihood ratio. Therefore, likelihood normalization is crucial in dealing with real-world data for speaker verification. Universal background model based normalization and cohort-based normalization are the two main approaches [17], [25]. The universal background model is a large speaker-independent GMM trained by pooling a great deal of speech data with the expectation–maximization (EM) algorithm [17]. Cohort-based normalization, which is very easy to implement, uses a set of cohort speakers that are close to the target speaker. The cohort can be seen as a replacement for the universal background model by calculating the probability of the cohort under the given conditions [17]. The essence of this paper is to investigate whether or not phase information is effective for robust speaker recognition. The consideration of state-of-the-art score normalization methods is, therefore, beyond the scope of this paper. Since it is very easy to implement, we use cohort-based normalization with the size of cohort set to 3 [30], [38].

IV. Experiments

A. Experimental Setup

NTT database [19] and a large-scale Japanese Newspaper Article Sentences (JNAS) database [27] were used to evaluate our proposed method.

NTT database is a standard database for Japanese speaker recognition. It has been used for many studies for speaker recognition [4], [18], [19], [30]–[34]. The NTT database consists of recordings of 35 (22 male and 13 female) speakers collected in five sessions over ten months (1990.8, 1990.9, 1990.12, 1991.3, and 1991.6) in a sound proof room. For training the models, we used the same five sentences (about 20 seconds of speech) for all speakers from one session (1990.8). These sentences were uttered in a normal speaking mode. Five other sentences from the other four sessions uttered at normal, fast and slow speeds were also used as test data. In other words, the test corpus consisted of 2100 trials for speaker identification, and 2100 true trials and 71 400 false trials for speaker verification. The average duration of a sentence is about 4 seconds. For the phase information, GMMs having 64 mixtures[4] with diagonal covariance matrices trained on five sentences were used as speaker models. For the MFCCs, the same five training sentences were also used to train the speaker-specific GMM for each speaker. We used a GMM with eight mixtures with full-covariance matrices which achieved the best speaker identification for various numbers of mixtures with full-covariance or diagonal-covariance matrices for the same NTT database [18]. With the popular NIST Speaker Recognition Evaluation (SRE) database [35], [36], GMMs having 512–2048 mixtures are normally used for speaker recognition. However, with the NTT database, almost all studies have reported that the best performances were achieved using a MFCC-based GMM having either 64 mixtures with diagonal covariance matrices or eight mixtures with full covariance matrices [4], [18], [19]. We guess that the reason for this is that the variations in speaking style (i.e., read speech) and microphone conditions in the NTT database are not very large.

For the JNAS database, 270 speakers (135 males and 135 females) in the JNAS database were used for speaker identification. Sentences from Japanese newspaper articles make up the content of utterances in the JNAS database. Sets of reading texts comprised 150 sets consisting of about 100 sentences each. Each speaker read one of the 150 sets. All utterances were collected with headset microphone. Five sentences (about 2 seconds/sentence[5]/sentence) were used for training speaker-special GMMs, and 95 sentences (about 5.5 seconds/sentence[5]5) were used for test. The number of mixtures for the phase-based GMMs and MFCC-based GMMs was both set to 64.

For both of the NTT database and JNAS database, the input speech was sampled at 16 kHz. 12 MFCCs and their first-order derivatives plus the first derivative of the power component (25 dimensions)[6] were calculated every 10 ms using a window of 25 ms. The phase information was calculated every 5 ms using a window of 12.5 ms[7]. The phase information was extracted as follows. First, a DFT for 256 sampling points was carried out, resulting in 128 individual components from 256 symmetrical components. To reduce the number of feature parameters, we used only phase information in a sub-band frequency range. The phase information was obtained from the lowest 12 components of the sub-band spectrum, corresponding to line spectra from 8000/128 Hz to $(8000/128) \times 12$ Hz (roughly speaking, from about 60 to 700 Hz). The interval between two adjacent components corresponds to about 60 Hz.

TABLE II
SPEAKER IDENTIFICATION RESULTS BY INDIVIDUAL METHOD ON NTT DATABASE (%)

speed		normal	fast	slow	Avg.
MFCC-based GMM		98.7	96.7	96.9	97.4
$\{\tilde{\theta}\}$	(60–700 Hz)	52.6	51.6	51.7	52.0
	(300–1000 Hz)	61.0	57.6	56.6	58.4
	(600–1300 Hz)	31.6	31.7	34.7	32.7
$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	(60–700 Hz)	73.4	72.0	70.4	71.9
	(300–1000 Hz)	60.4	55.6	57.6	57.9
	(600–1300 Hz)	37.4	37.1	39.0	37.8
spectrum	(60–700 Hz)	76.4	73.7	74.9	75.6
	(300–1000 Hz)	67.9	68.0	65.3	67.1
	(600–1300 Hz)	60.9	61.1	61.6	61.2
	(1060–1760 Hz)	56.1	51.0	55.6	54.2
	(1480–2180 Hz)	49.3	47.0	46.9	47.7

B. Speaker Identification Results

1) Speaker Identification on NTT Database: We evaluated the text-independent speaker identification experiment using the phase information on NTT database in this section.

The speaker identification results by individual method are shown in Table II. The method phase $\{\tilde{\theta}\}$ means that the phase value obtained by (6) was used as the speaker identification feature. The method phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ means that the phase values $\{\tilde{\theta}\}$ are transformed to coordinates by (7), resulting in double the number of parameters compared with phase $\{\tilde{\theta}\}$. The modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ significantly outperformed the original phase $\{\tilde{\theta}\}$ [28] because the difference of the modified phase is more correctly modeled by GMMs than that of the original phase. For the modified phase information, the first 12 feature parameters, that is, from the first component to twelfth component of the spectrum (frequency range: 60–700 Hz) achieved the best identification performance of all the other sub-band frequency ranges. Although the phase information-based method performed worse than the MFCC-based method, it is useful for speaker recognition. In comparison, we also used the logarithmic power spectrum in a sub-band frequency range obtained from (1), that is, $\log(X^2(\omega,t) + Y^2(\omega,t))$. In [29], Matsui and Tanabe reported that the logarithmic power spectrum is also effective for speaker recognition. For individual methods, the best performance of the modified phase information $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ (60–700 Hz) was similar to that of the sub-band frequency range of the logarithmic power spectrum.

Fig. 5 shows the speaker identification results using a combination of MFCCs and the logarithmic power spectrum in a sub-band frequency range. The speaker identification results for combining MFCCs with the original and modified phase information are shown in Figs. 6 and 7, respectively. All the combina-

[4] GMMs with different number of mixtures are also used for speaker identification; the combination of the phase-based GMMs with 64 mixtures and the MFCC-based method achieved the best performance. Due to limited space, the results for GMMs with different number of mixtures were not described.

[5] excluding about 2 second silence at beginning and endding of a sentence

[6] DFT with 512 samples (400 points of data plus 112 zeros) was used.

[7] DFT with 256 samples (200 points of data plus 56 zeros) was used.

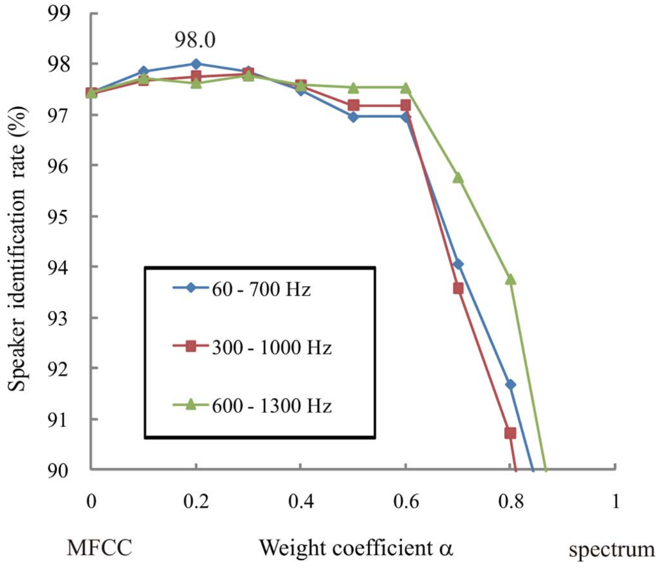


Fig. 5. Speaker identification results using a combination of MFCC and log-power spectrum on NTT database (average of three speaking modes).

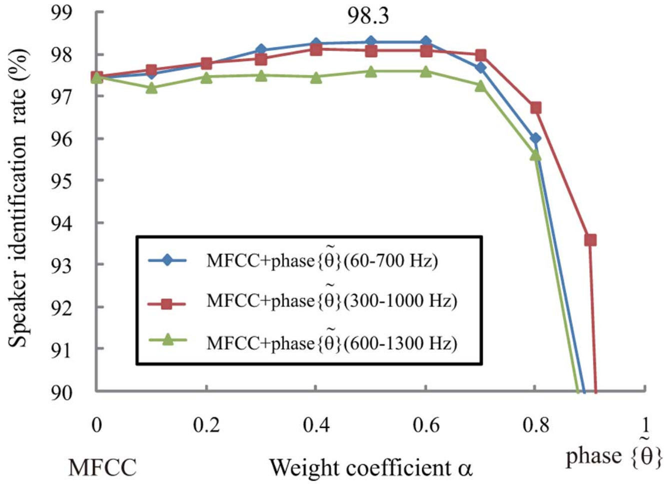


Fig. 6. Speaker identification results using a combination of MFCC and the original phase $\{\tilde{\theta}\}$ on NTT database (average of three speaking modes).

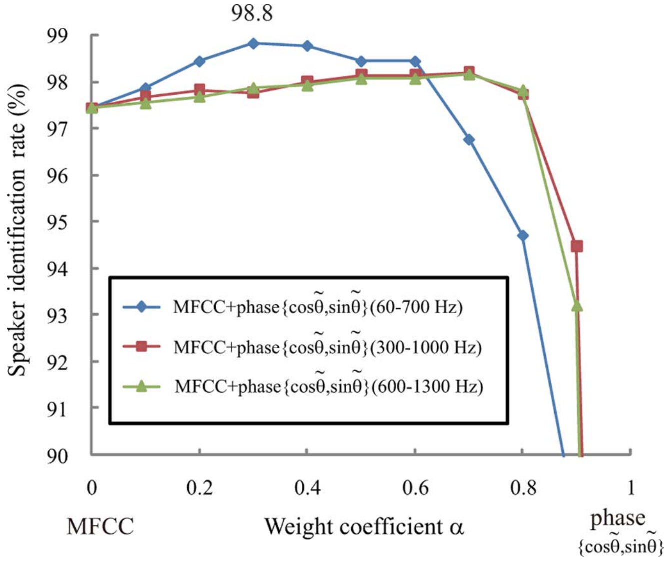


Fig. 7. Speaker identification results using the combination of MFCC and the modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ on NTT database (average of three speaking modes).

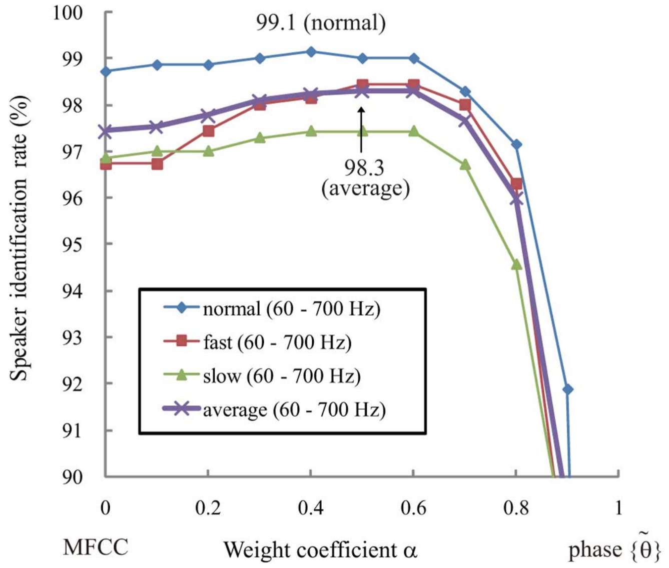


Fig. 8. Speaker identification results using the combination of MFCC and the original phase $\{\tilde{\theta}\}$ on NTT database (frequency range: 60–700 Hz).

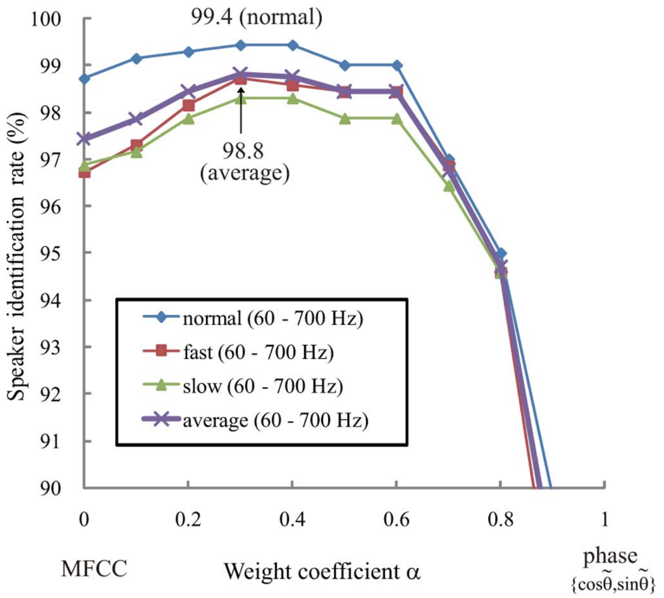


Fig. 9. Speaker identification results using a combination of MFCC and the modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ on NTT database (frequency range: 60–700 Hz).

tion methods improved the speaker identification performance significantly compared with the individual MFCC-based GMM. The best performances of the combination of MFCCs with various sub-band frequency ranges of the log-power spectrum are similar (about 98.0%). Both the original and modified phase information methods achieved best performances (98.3%, and 98.8%, respectively) in the 60–700 Hz frequency range when combined with MFCCs. MFCC has strong correlation with the log power spectrum (which is also influenced by the harmonics of fundamental frequency) while it has poor correlation with the phase information. The combination of MFCC and the phase information performed better than the combination of MFCC and the log power spectrum indicates that the phase information has high complement with MFCC and it shows the effectiveness for speaker recognition. When phase information was combined

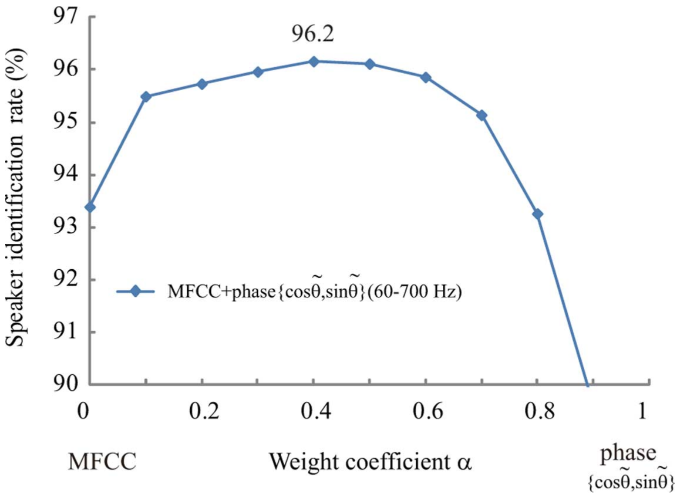


Fig. 10. Speaker identification results using the combination of MFCC-based GMM and the modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ on JNAS database.

with MFCCs, the modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ performed significantly better than the original phase $\{\tilde{\theta}\}$. The phase components derived from the 300–1000 Hz sub-band, which belongs to the frequency range of telephone channels, are also very effective for speaker identification. We also compared the speaker identification performances of the original and modified phase information for all speaking modes in the 60–700 Hz frequency range as illustrated in Figs. 8 and 9, respectively. The proposed modified phase information $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ is more robust than the proposed original phase information $\{\tilde{\theta}\}$ for all speaking modes. The combination of MFCCs and the modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ achieved a relative error reduction rate of 53.8% over the MFCC-based method for normal speed utterances, 60.6% reduction for fast speed and 45.2% reduction for slow speed.

2) Speaker Identification on a Large-Scale JNAS Database: The experiments of Section IV-BI show that the combination of the MFCC-based GMM and the modified phase-based GMM is significantly better than individual MFCC-based GMM. For the NTT database, although it includes four different recording sessions and three speaking modes for the test, the test trials $(= 2100)$ for speaker identification are relatively small, To verify the robustness of the proposed method for speaker identification, we also conducted the modified phase-based method on a large scale JNAS database [27]. For the large-scale JNAS database, the test trials are 25 650 (270 × 95).

Fig. 10 shows the speaker identification results on JNAS database. The speaker identification rates of MFCC-based GMMs and modified phase-based GMMS were 93.4% and 73.5%, respectively. The result of combination of MFCC-based GMMs and modified phase-based GMMs was improved to 96.2% (the error reduction rate of 42.4%). The experimental results show that the combination method improved the speaker recognition performance significantly on a large-scale speech corpus than the MFCC-based method.

C. Speaker Verification Results

The effectiveness of using the phase information for speaker identification was demonstrated in Section IV-B. The proposed phase information (frequency range: 60–700 Hz) was also used to perform speaker verification[8] in this section.

The speaker modeling techniques used for speaker identification were also used for speaker verification [2], [5], [25]. The experimental setup, as well as the speech analysis conditions, were also the same as those for speaker identification.

TABLE III
EQUAL ERROR RATE FOR SPEAKER VERIFICATION ON NTT DATABASE (%)

speed	normal	fast	slow	Avg.
MFCC	0.42	0.85	0.90	0.72
spectrum (60–700 Hz)	6.67	7.94	7.21	7.27
$\{\tilde{\theta}\}$	9.62	9.86	9.58	9.69
$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	5.16	6.15	5.69	5.66
MFCC+$\{\tilde{\theta}\}$	0.34	0.67	0.80	0.60
MFCC+$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	0.28	0.51	0.55	0.45
MFCC+spectrum (60–700 Hz)	0.38	0.78	0.78	0.65

TABLE IV
EQUAL ERROR RATE FOR SPEAKER VERIFICATION ON JNAS DATABASE (%)

MFCC	$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	MFCC + $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$
1.24	2.61	0.93

1) Speaker Verification on NTT Database: The equal error rate (EER) for speaker verification on NTT database is given in Table III. The trend in the speaker verification results is similar to that in the speaker identification results. Although the phase information performed worse than the MFCCs, a relatively high verification performance (about 5.7% ERR for the modified phase information $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$) was obtained. The modified phase information $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ improved the speaker verification performance significantly over the original phase information $\{\tilde{\theta}\}$ for all cases. The logarithmic power spectrum in a sub-band frequency range (60–700 Hz) was compared with the original/modified phase information. Unlike the speaker identification, the performance of the logarithmic power spectrum was between the original phase information and the modified phase information for speaker verification with an individual method. The modified phase information $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ significantly outperformed the log-power spectrum for all cases. Because of the high complement of the MFCCs and the phase information, the combination of MFCCs and phase improved the speaker verification performance remarkably. The combination of the MFCC and the modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ achieved a relative error reduction rate of 38.9% over the MFCC on NTT database.

2) Speaker Verification on JNAS Database: The speaker verification EER on JNAS database is given in Table IV. The combination of the MFCC and the modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ achieved a relative error reduction rate of 25.0% over the MFCC on JNAS database.

[8] 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 [1].

TABLE V
Statistical Test (Sign Test) for Speaker Identification on NTT Database by Combining MFCC and Phase/Spectrum. The '$+$' Sign Indicates the Combination of two Type of Feature Set. Plus Sign for Sign Test Means the Hypothesis "$A < B$" Is True

hypothesis	significance level (%)	numbers of (plus sign, minus sign) the sign test
MFCC< MFCC+spectrum	3.92	(16,5)
MFCC $<$ MFCC+$\{\tilde{\theta}\}$	0.10	(22,4)
MFCC $<$ MFCC+$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	< 0.01	(32,3)
(MFCC+spectrum) $<$ MFCC+$\{\tilde{\theta}\}$	> 10	(16,9)
(MFCC+spectrum) $<$ MFCC+$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	< 0.01	(9,1)
(MFCC+$\{\tilde{\theta}\}$) $<$ MFCC+$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	2.20	(15,4)

TABLE VI
Statistical Test (Sign Test) for Speaker Verification on NTT Database by Combining MFCC and Phase

hypothesis	significance level (%)	numbers of (plus sign, minus sign) the sign test
MFCC $<$ MFCC+$\{\tilde{\theta}\}$	< 0.01	(163,92)
MFCC $<$ MFCC+$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	< 0.01	(287,125)
(MFCC+$\{\tilde{\theta}\}$) $<$ MFCC+$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	< 0.01	(193,102)

D. Significance Test Results

Since the scale of the NTT database is relatively small, the sign test[9] was performed on NTT database to evaluate the "significance difference" of different methods.

Statistical test (sign test) results for speaker identification by combining the MFCC and phase information/spectrum are shown in Table V. Speaker identification using a combination of the MFCC and the modified phase information $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ performed significantly better than that using either the MFCC or a combination of MFCC and spectrum at a significance level at a significance level[10] less than 0.01%. It also performed better than that using a combination of MFCC and the original phase information $\{\tilde{\theta}\}$ at a significance level less than 5% ($\approx 2.2\%$). However, speaker identification using a combination of MFCC and the original phase information $\{\tilde{\theta}\}$ was not significantly better than that using a combination of MFCC and the spectrum at a significance level less than 10%.

Statistical test (sign test) results for speaker verification by combining the MFCC and modified phase information are shown in Table VI. Speaker verification using a combination of the MFCC and the phase information $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ performed significantly better than that using the MFCC and the combination of the MFCC and the original phase information $\{\tilde{\theta}\}$ at a significance level less than 0.01%, respectively.

V. Related Work

We now compare our method with related research using the same NTT database; that is, with 22 male and 13 female speakers. Matsui and Furui reported speaker identification rates of 95.1%, 91.5%, and 93.1% for utterances at normal, fast, and slow speaking rates, respectively. They used GMMs having 64 mixtures with diagonal covariance matrices trained with ten utterances [19].

Markov and Nakagawa reported speaker identification rates of 97.9%, 94.1%, and 95.7%, and speaker verification EERs of 0.16%, 0.86%, and 0.82% for utterances at normal, fast and slow speaking rates, respectively [30]. They used GMMs having eight mixtures with full covariance matrices trained with ten utterances and a nonlinear frame likelihood transformation. They also evaluated their methods on a large-scale TIMIT database [37], containing a total of 6300 sentences, ten sentences spoken by each of 630 speakers from eight major dialect regions of the United States. Their methods also achieved a high speaker identification rate of 99.6% and speaker verification EER of 0.01%. Thus, it is possible that our proposed method will also be effective for large-scale speaker databases such as the TIMIT database. Although our proposed combination of MFCCs and phase information used only five utterances to train the speaker model, it nevertheless significantly outperformed the methods of Markov and Nakagawa. They also proposed a GMM-based text independent speaker recognition system integrating the pitch and LPC residual with the LPC derived cepstral coefficients [38]. This system achieved speaker identification rates of 98.1%, 97.3%, and 96.1%, and speaker verification EERs of 0.29%, 1.18%, and 1.19% for utterances at normal, fast, and slow speaking rates, respectively. They used GMMs having 64 mixtures with diagonal covariance matrices trained by ten utterances [38].

Miyajima *et al.* reported an identification rate of 99.0% for utterances at a normal speaking rate [31]. They used GMMs trained by 15 utterances, integrating cepstral coefficients and

[9] A significance test is performed to determine if an observed value of a statistic differs enough from a hypothesized value of a parameter to draw the inference that the hypothesized value of the parameter is not the true value.

In statistics, the sign test is one of the significance test method which is a non parametric test. The sign test is used to test the null hypothesis and whether or not two groups are equally sized. Sign test is based on the direction of the plus and minus sign of the observation for matched pairs $\{x_i, y_i\}$, but not their numerical magnitude.

[10] The significance level of a test is a traditional frequentist statistical hypothesis testing concept. In simple cases, it is defined as the probability of making a decision to reject the null hypothesis when the null hypothesis is actually true (a decision known as a Type I error, or "false positive determination"). Popular levels of significance are 5%, 1%, and 0.1%. For example, if someone argues that "there's only one chance in a thousand this could have happened by coincidence," a 0.1% level of statistical significance is being implied.

pitch, and estimated by the MCE. Miyajima's group also proposed a parameter sharing method in Gaussian mixtures based on a mixture of factor analysis and a discriminative training method [32]. They showed identification rates from about 97.3% to about 98.6% under the above experimental conditions.

Nishida and Ariki reported an identification rate of 94.9% for utterances at a normal speaking rate [33]. They used a subspace method that maps speech separately to a phonetic space and a speaker characteristic space. The speaker model was trained by five utterances.

Our proposed method significantly outperforms the methods discussed in the related research section above.

VI. Conclusion

In this paper, we proposed an original and modified phase information extraction method that normalizes the change variation in the phase according to the frame position of the input speech and combined the phase information with MFCCs in speaker identification and verification methods. The speaker identification and verification experiments were conducted on the NTT database which consists of sentences data uttered at normal, slow and fast speaking modes by 35 Japanese speakers. The proposed modified phase information $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ remarkably improved the identification performance for all speaking modes compared with the original phase information $\{\tilde{\theta}\}$. Combining the MFCC and the modified phase information, we obtained the error reduction rate of 53.8%, 60.6%, and 45.2% over MFCC-based method for normal, fast and slow speaking modes, respectively. These results show the best performance in comparison with the other rescarchers' results for the same database [4], [18], [19], [30]–[32], [38]. To verify the robustness of phase information for speaker recognition, the modified phase information $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ was also used to perform speaker verification. The experiments of the combination showed the speaker verification EER of 0.28% for normal, 0.51% for fast and 0.55% for slow speaking modes, respectively. By integrating the MFCC with the modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$, a relative error reduction rate of 38.9% from MFCC was achieved.

We also conducted the speaker identification and verification experiments on the large scale JNAS database, a similar trend as the NTT database was obtained.

The speaker identification by phase information for noisy speech was also evaluated in [39]. The experimental results showed that the combination of the phase information and MFCC was also very effective for noisy speech. The speaker recognition rates of 76.3% for MFCC, 64.7% for the phase information and 91.6% for the combination of MFCC and the phase information were obtained under 20-dB conditions with stational/non-stational noise using speaker models trained by clean speech [39]. In our future work, we intend investigating the performance of the proposed phase feature under various channels and background conditions.

References

[1] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, no. 4, pp. 475–487, Apr. 1976.

[2] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, no. 1–2, pp. 91–108, 1995.

[3] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[4] S. Nakagawa, W. Zhang, and M. Takahashi, "Text-independent speaker recognition by combining speaker specific GMM with speaker adapted syllable-based HMM," in *Proc. ICASSP*, 2004, vol. I, pp. 81–84.

[5] F. Bimbot *et al.*, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, pp. 430–451, 2004.

[6] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Commun.*, vol. 22, pp. 403–417, 1997.

[7] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. Eurospeech'03*, 2003, pp. 2117–2120.

[8] G. Shi *et al.*, "On the importance of phase in human speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1867–1874, Sep. 2006.

[9] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," *Proc. ICASSP*, vol. 1, pp. 133–136, 2001.

[10] P. Aarabi *et al., Phase-Based Speech Processing*. Singapore: World Scientific, 2005.

[11] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, Jan. 2006.

[12] R. M. Hegde, H. A. Murthy, and G. V. R. Rao, "Application of the modified group delay function to speaker identification and discrimination," in *Proc. ICASSP*, 2004, pp. 517–520.

[13] R. Padmanabhan, S. Parthasarathi, and H. Murthy, "Robustness of phase based features for speaker recognition," in *Proc. Interspeech*, 2009, pp. 2355–2358.

[14] J. Kua, J. Epps, E. Ambikairajah, and E. Choi, "LS regularization of group delay features for speaker recognition," in *Proc. Interspeech*, 2009, pp. 2887–2890.

[15] [Online]. Available: http://www.nist.gov/speech/tests/sre/

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

[17] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Dig. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[18] S. Nakagawa, W. Zhang, and 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.

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

[20] [Online]. Available: http://speechlab.bu.edu/VTCalcs.php

[21] S. Young, D. Kershow, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 3.0)*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[22] J. Gauvain and C. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov Chains," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[23] Y. Tsurumi and S. Nakagawa, "An unsupervised speaker adaptation method for continuous parameter HMM by maximum *a posteriori* probability estimation," in *Proc. ICSLP'94*, 1994, pp. 431–434.

[24] N. Nilsson, *Learning Machines*. New York: McGraw-Hill, 1966.

[25] 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. Int. Conf. Spoken Lang. Processing (ICSLP '92)*, 1992, vol. 1, pp. 599–602.

[26] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu, "Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition," in *Proc. Int. Workshop Autom. Speech Recognit. Understanding*, 1999, pp. 393–396.

[27] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *J. Acoust. Soc. Jpn. (E)*, vol. 20, no. 3, pp. 199–206, 1999.

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

[29] T. Matusi and K. Tanabe, "Comparative study of speaker identification methods: DPLRM, SVM and GMM," in *Proc. IEICE, E89-D(3)*, 2006, pp. 1066–1073.

[30] K. P. Markov and S. Nakagawa, "Text-independent speaker recognition using non-linear frame likelihood transformation," *Speech Commun.*, vol. 24, no. 3, pp. 193–209, 1998.

[31] C. Miyajima, Y. Hattori, and K. Tokuda, "Text-independent speaker identification using Gaussian mixture models based on multi-space probability distribution," *IEICE Trans.*, vol. E84-D, no. 7, pp. 847–855, 2001.

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

[33] M. Nishida and Y. Ariki, "Speaker recognition by separating phonetic space and speaker space," in *Proc. Eurospeech*, 2001, pp. 1381–1384.

[34] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," *Speech Commun.*, vol. 49, no. 6, pp. 501–513, Jun. 2007.

[35] [Online]. Available: http://www.nist.gov/speech/tests/sre/

[36] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation-Overview, methodology, systems, results, perspective," *Speech Commun.*, vol. 31, no. 2–3, pp. 225–254, 2000.

[37] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, vol. 9, no. 4, pp. 351–356, 1990.

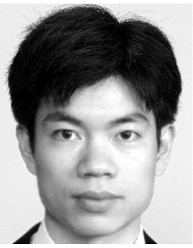
[38] K. P. Markov and S. Nakagawa, "Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition," *J. ASJ (E)*, vol. 20, no. 4, pp. 281–291, 1999.

[39] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining MFCC and phase information in noisy environments," in *Proc. ICASSP*, 2010, pp. 4502–4505.

Seiichi Nakagawa (M'87) received the B.E. and M.E. degrees from the Kyoto Institute of Technology, Kyoto, Japan, in 1971 and 1973, respectively, and the Dr. of Eng. degree from Kyoto University in 1977.

He joined the faculty of Kyoto University in 1976 as a Research Associate in the Department of Information Sciences. From 1980 to 1983, he was an Assistant Professor, and from 1983 to 1990 he was an Associate Professor. Since 1990 he has been a Professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi, Japan. From 1985 to 1986, he was a Visiting Scientist in the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA. His major interests in research include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence.

Prof. Nakagawa received the 1997/2001 Paper Award from the IEICE and the 1988 JC Bose Memorial Award from the Institution of Electronics and Telecommunications Engineers. He is a fellow of the Institute of Electronics, Information, and Communication Engineers (IEICE) and the Information Processing Society of Japan (IPSJ).



Longbiao Wang (M'09) received the B.E. degree from Fuzhou University, Fuzhou, China, in 2000 and the M.E. and Dr. Eng. degrees from Toyohashi University of Technology, Toyohashi, Japan, in 2005 and 2008, respectively.

From July 2000 to August 2002, he worked at the China Construction Bank. Since 2008 he has been an Assistant Professor in the faculty of Engineering, Shizuoka University, Shizuoka, Japan. His research interests include robust speech recognition, speaker recognition and sound source localization.

He received the "Chinese Government Award for Outstanding Self-financed Students Abroad" in 2008. He is a member of the IEICE and the Acoustical Society of Japan (ASJ).

Shinji Ohtsuka received the B.E. degree from Toyohashi University of Technology, Toyohashi, Japan, in 2005.

His research interests include speaker recognition. He is currently with Yazaki Syscomplus Co., Ltd, Makinohara, Japan.