PAPER Special Section on Processing Natural Speech Variability for Improved Verbal Human-Computer Interaction

Speaker Recognition by Combining MFCC and Phase Information in Noisy Conditions

Longbiao WANG[†a)], ***Member***, **Kazue MINAMI**[††], ***Nonmember***, **Kazumasa YAMAMOTO**[††], ***Member***, ***and*** **Seiichi NAKAGAWA**[††], ***Fellow***

SUMMARY In this paper, we investigate the effectiveness of phase for speaker recognition in noisy conditions and combine the phase information with mel-frequency cepstral coefficients (MFCCs). To date, almost speaker recognition methods are based on MFCCs even in noisy conditions. For MFCCs which dominantly capture vocal tract information, only the magnitude of the Fourier Transform of time-domain speech frames is used and phase information has been ignored. High complement of the phase information and MFCCs is expected because the phase information includes rich voice source information. Furthermore, some researches have reported that phase based feature was robust to noise. In our previous study, a phase information extraction method that normalizes the change variation in the phase depending on the clipping position of the input speech was proposed, and the performance of the combination of the phase information and MFCCs was remarkably better than that of MFCCs. In this paper, we evaluate the robustness of the proposed phase information for speaker identification in noisy conditions. Spectral subtraction, a method skipping frames with low energy/Signal-to-Noise (SN) and noisy speech training models are used to analyze the effect of the phase information and MFCCs in noisy conditions. The NTT database and the JNAS (Japanese Newspaper Article Sentences) database added with stationary/non-stationary noise were used to evaluate our proposed method. MFCCs outperformed the phase information for clean speech. On the other hand, the degradation of the phase information was significantly smaller than that of MFCCs for noisy speech. The individual result of the phase information was even better than that of MFCCs in many cases by clean speech training models. By deleting unreliable frames (frames having low energy/SN), the speaker identification performance was improved significantly. By integrating the phase information with MFCCs, the speaker identification error reduction rate was about 30%–60% compared with the standard MFCC-based method.

key words: *speaker identification, phase information, MFCC, noisy environment, GMM*

1. Introduction

Speaker recognition performance degrades remarkably in noisy environments [1]–[9]. Missing feature theory and spectral subtraction were used in many speaker recognition task in noisy environments [1]–[4]. In [2], a method that combined multicondition model training and missing-feature theory to model noise with temporal-spectral characteristics was proposed. Multicondition training was conducted using simulated noisy data of simple noise characteristics, providing a coarse compensation for the noise, and missing-feature theory was applied to refine the compensation by ignoring noise variation outside the given training conditions, thereby accommodating training and testing mismatch. The speaker identification rate of the proposed method was improved to 85%–94% from 53%–82% for 20 dB noisy speech and improved to 51%–75% from 12%–43% for 10 dB noisy speech by MFCC-based GMMs trained on the multicondition data for TIMIT database [2]. Parallel model combination (PMC) technique was also applied in speaker recognition [5], [6]. The use of microphone arrays to improve noise robustness for speaker recognition has been discussed in [7], [8]. A noise-robust multi-stream speaker verification method using F_0 information was proposed and evaluated in [9].

Almost all of above methods were based MFCCs, which only used the magnitude of the Fourier Transform of time-domain speech frame, that is, the phase component has been ignored. The MFCCs dominantly capture the speaker-specific vocal tract information. Feature parameters extracted from excitation source characteristics are also useful for speaker recognition [10]–[13]. Almost all of these are based on Linear Predictive Coding (LPC) analysis. Therefore, the separation between sound source characteristics and vocal tract characteristics is incomplete. Markov and Nakagawa proposed a GMM-based text-independent speaker recognition system integrating the pitch and LPC residual with the LPC derived cepstral coefficients [10]. Their experimental results showed that using pitch information was most effective when the correlation between pitch and the cepstral coefficients was taken into consideration. Zheng et al. proposed a speaker verification system using complementary acoustic features derived from the vocal source excitation and the vocal tract system [12], [13].

The importance of phase in human speech recognition has been reported in [14]–[17]. Paliwal and Alsteris also investigated the relative importance of short-time magnitude and phase spectra on speech perception [14]. Human perception experiments were conducted to measure intelligibility of speech tokens synthesized from either the magnitude or phase spectrum. It was shown in [14] 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 [15], 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 indicated that a small amount of phase er-

Manuscript received December 3, 2009.
Manuscript revised March 1, 2010.

[†]The author is with Shizuoka University, Hamamatsu-shi, 432–8560 Japan.
[††]The authors are with Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.
a) E-mail: wang@sys.eng.shizuoka.ac.jp

DOI: 10.1587/transinf.E93.D.2397

ror 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, we expect that the phase may also be important in automatic speech/speaker recognition. In particular, the phase is important for speaker recognition, because it can convey voice source information.

Recently, many speaker recognition studies using group delay based phase information have been proposed [18]–[20]. Group delay is defined as the negative derivative of the phase of the Fourier transform of a signal. So it is very related to our phase information. In [19], 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 [21] resulted in better performance for modified group delay features [18] (about 15% EER) when compared to conventional MFCC features (about 18% EER). In [20], the authors proposed an alternative complementary feature extraction method to reduce the variability of group delay features derived from the speech spectrum with least squares regularization. The proposed log compressed least square group delay achieved 10.01% Equal Error Rate (EER) compared to 7.64% EER for MFCC, and the fusion of group delay and MFCC improved to 7.16% EER for NIST 2001 SRE database. Evaluations on the NIST 2008 SRE databases showed a relative improvement of 18% EER respectively when group delay-based system was fused with MFCC-based system. Actually, the group delay based phase contains both the power spectrum and phase information, so the complementary nature of power spectrum-based MFCC and group delay phase was not remarkable.

In this paper, we focus on the investigation whether or not the phase information without power-spectrum information directly extracted from Discrete Fourier Transform (DFT) of an input speech is effective for speaker recognition in noisy conditions. In our previous study [22], [23], we proposed a phase information extraction method that normalizes the change variation in phase $\{\tilde{\theta}\}$ or $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ depending on the clipping position of the input speech, even with the same frequency.

The proposed phase information was very effective for speaker identification and speaker verification for clean speech [22], [23]. In this paper, we investigate the robustness of the proposed phase information for speaker identification in noisy conditions and combine the phase information with MFCCs. Various speech databases added with various noise nature and Signal-to-Noise Ratio (SNR) are used to verify the robustness of phase to noise. Various noise-robust processing techniques, such as a special missing feature theory which skips frames with low energy/SN, spectral subtraction and noisy speech training models etc. are used for the phase information based method or MFCC-based method, which aim to analyze the effect of the phase information and MFCCs under match and mismatch conditions in noisy environments. To verify the robustness of the phase-based method for speaker recognition in noisy environments, we conducted the speaker recognition experiments on a small scale NTT database [30] and a large scale JNAS (Japanese Newspaper Article Sentences) database [31].

The remainder of this paper is organized as follows: Sect. 2 formulates the phase information. Section 3 briefly describes the combination method for speaker recognition. Some noise-robust techniques, such as noisy speech training models, spectral subtraction and a special missing feature theory are introduced in Sect. 4. The experiments for speaker identification using phase information in noisy conditions are evaluated in Sect. 5. Finally, Sect. 6 summarizes this paper.

2. Phase Information Extraction [22], [23]

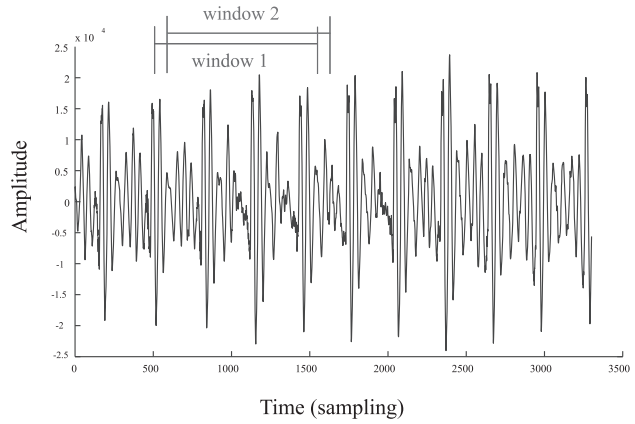
In [22], [23], we investigated the effect of phase on speaker recognition using both synthesized and human speech. The conclusion reached was that phase information was effective for speaker recognition. In this section, a phase information extraction method is described.

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

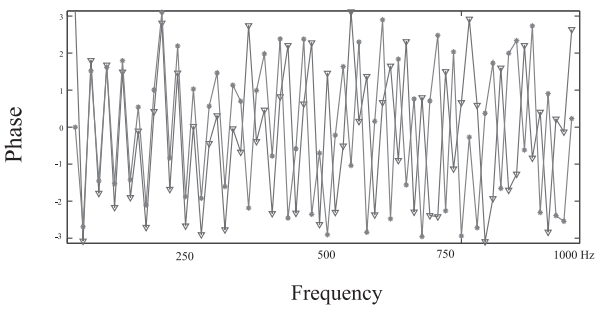
$$\begin{aligned} S(\omega, t) &= X(\omega, t) + jY(\omega, t) \\ &= \sqrt{X^2(\omega, t) + Y^2(\omega, t)} \times e^{j\theta(\omega, t)}, \end{aligned} \tag{1}$$

where ω is radian frequency, $X(\omega, t)$ and $Y(\omega, t)$ are the real part and imaginary part of spectrum, respectively. For conventional MFCCs, 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 parameters for speaker recognition. GMMs used in this paper are insensitive to the temporal aspects of the speech. Thus, $\theta(\omega, t)$ and $2\pi + \theta(\omega, t)$ can express the same speaker characteristics by GMMs by constraining phases extracted from all frames to $[0, 2\pi]$ or $[-\pi, \pi]$. In this paper, all the phases are constrained to $[-\pi, \pi]$.

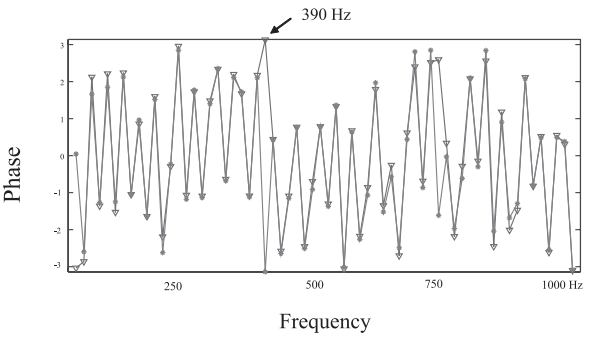
However, the phase $\theta(\omega, t)$ changes depending on the clipping position of the input speech even with the same frequency ω. To help the reader to understand the effectiveness of our proposed phase processing, an example of the effect of the clipping position on phase for Japanese vowel /a/ is illustrated in Fig. 1. As shown in Fig. 1 (b), the unnormalized wrapped phases of two windows become quite a bit different because the phases change depending on the clipping position. X-axis means frequency and y-axis means phase value from $-\pi$ to π. In Fig. 1 (b), horizontal axis is frequency in Hz and vertical axis is phase value. It is obvious that the phase $\theta(\omega, t)$ differs for different clipping positions. For speaker recognition using phase information, the phases extracted from two different windows of the same sentence from the same people should be as small as possible. Thus, it is necessary to normalize the phase distortion to decrease the difference of two phases from two different windows.



(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

Fig. 1 Example of the effect of clipping position on phase for Japanese vowel /a/.

A basic processing for the elimination of the different position influences is explained as following. Let $s_1, s_2, \cdots, s_{128}, \{s_{129} = s_1\}$ be the sampling sequence for a cyclic function. The phase of $s_1, s_2, \cdots, s_{128}$ and the phase of $s_2, \cdots, s_{128}, s_1$ are different each other. The difference for the radian frequency ω is $1/128 \times \omega/2\pi$. By using this relation, we can get the normalizing equation.

To overcome this problem, the phase of a certain basis radian frequency ω_b of all frames is converted to constant, and the phase of the other frequency is estimated relative to this. In the experiments discussed in this paper, the phase of basis radian frequency ω_b is set to $2\pi \times 1000$ Hz. For example, setting the phase of the basis radian frequency $\theta(\omega_b, t)$ to $\pi/4$, we have

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

The difference of unnormalized wrapped phase $\theta(\omega_b, t)$ on basis frequency ω_b in Eq. (1) and the normalized wrapped phase in Eq. (2) is $(\frac{\pi}{4} - \theta(\omega_b, t))$. With $\omega = 2\pi f$ in the other frequency (that is, $\omega \neq 2\pi \times 1000$ Hz), the difference becomes $\frac{\omega}{\omega_b}(\frac{\pi}{4} - \theta(\omega_b, t))$. Thus, the spectrum on frequency ω becomes

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

and the phase can be normalized. Then, the real and imaginary parts of Eq. (3) are given by

$$\tilde{X}(\omega, t) = \sqrt{X^2(\omega, t) + Y^2(\omega, t)} \times \cos\left\{\theta(\omega, t) + \frac{\omega}{\omega_b}\left(\frac{\pi}{4} - \theta(\omega_b, t)\right)\right\}, \quad (4)$$

$$\tilde{Y}(\omega, t) = \sqrt{X^2(\omega, t) + Y^2(\omega, t)} \times \sin\left\{\theta(\omega, t) + \frac{\omega}{\omega_b}\left(\frac{\pi}{4} - \theta(\omega_b, t)\right)\right\}, \quad (5)$$

and the phase information is normalized as

$$\tilde{\theta}(\omega, t) = \theta(\omega, t) + \frac{\omega}{\omega_b}\left(\frac{\pi}{4} - \theta(\omega_b, t)\right), \quad (6)$$

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

To reduce the number of feature parameters, we used only phase information in a sub-band frequency range. 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 changed the phase into coordinates on a unit circle, that is,

$$\tilde{\theta} \rightarrow \{\cos\tilde{\theta}, \sin\tilde{\theta}\}. \quad (7)$$

In this paper, the proposed modified phase information $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ is used to perform speaker identification in noisy conditions.

The unnormalized wrapped phase obtained from Eq. (1) and the normalized wrapped phase obtained from Eq. (6) were compared in Fig. 1 (b) and (c). After normalizing the wrapped phase by Eq. (6), the phase values shown in Fig. 1 (c) become very similar, that is to say, the normalized wrapped phase is more adaptable to speaker recognition. For this example, the Euclidian distance of unnormalized wrapped phases and normalized wrapped phases of two

different clipping 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 of two original unnormalized wrapped phases obtained from Eq. (6) were 6.2696 even 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 Eq. (7), the difference became to $\{0, 0.01\}$. Thus, the modified normalized wrapped phase was more effectively capture the speaker characteristics.

3. Combination Method

A Gaussian Mixture Model (GMM) is widely used as a speaker model [24]–[27]. 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, 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 log 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 [26], [27]

$$L^n_2 = (1-\beta)L^n_{MFCC} + \beta L^n_{phase}, \quad n = 1, 2, \cdots, N, \quad (8)$$

where L^n_{MFCC} and L^n_{phase} are the log likelihood produced by the *n-th* MFCC-based speaker model and phase information based speaker model, respectively. N is the number of speakers registered and β denote weighting coefficients. A speaker with maximum score is decided as the target speaker.

4. Speaker Identification Methods in Noisy Environment

Speaker identification performance degrades remarkably in noisy environments [2], [3]. The standard approches are based on multicondition model training, missing theory [2], [3] or spectral subtraction. In this paper, we focus on the investigation whether or not the phase information is effective for speaker identification in noisy environments. We use models trained by noisy data, a special technique of missing feature theory, and spectral subtraction.

4.1 Spectrum Subtraction with Smoothing of Time Direction

The observation signal x is assumed to be the sum of speech signal s and noise n, namely, $x = s + n$. Spectral subtraction [28] in the power spectral domain is defined as below:

$$|\tilde{S}_i(t)|^2 = |X_i(t)|^2 - \alpha|\tilde{N}_i|^2, \quad (9)$$

where $|\tilde{S}_i(t)|^2$ and $|X_i(t)|^2$ are the i-th components of the estimated power spectrum of clean speech and the power spectrum of observed signals at time t, respectively, while $|\tilde{N}_i|^2$ is the i-th component of *apriori* estimated power spectrum of noise, and α is the overestimation factor. In this paper, the i-th components of the estimated spectrum of clean speech at time t $\tilde{S}_i(t)$ which is calculated from the estimated power spectrum of clean speech $|\tilde{S}_i(t)|^2$ and phase of observed noisy speech at time t, and it is used to reconstruct the clean speech. We can express $|X_i(t)|^2$ as:

$$|X_i(t)|^2 = |S_i(t)|^2 + |N_i(t)|^2 + 2|S_i(t)||N_i(t)|\cos\theta_i(t), \quad (10)$$

where $|S_i(t)|$ and $|N_i(t)|$ are the true values for speech and noise, and $\theta_i(t)$ is the phase difference between speech and noise.

Here, we define the smoothing method of time direction to eliminate the effect of the phase difference between speech and noise as follows [29]:

$$\overline{|X_i(t)|^2} \approx |S_i(t)|^2 + |N_i(t)|^2. \quad (11)$$

Replacing $|X_i(t)|^2$ in Eq. (9) with $\overline{|X_i(t)|^2}$, Eq. (9) becomes

$$|\tilde{S}_i(t)|^2 \approx |S_i(t)|^2 + |N_i(t)|^2 - \alpha|\tilde{N}_i|^2. \quad (12)$$

Therefore, we can estimate the speech signal more accurately if we can estimate $|\tilde{N}_i|$ accurately. Here, $|\tilde{N}_i|$ was estimated by averaging power spectrum of silence speech (noise only) at beginning of test sentence.

4.2 Skipping Low Energy/SN Frames or Unreliable Frames

Even if noise is stationary, SN ratio at every frame depends on the power of speech. For example, the power of vowel is stronger than that of unvoiced consonants, in other words, the SN ratio for vowel parts is larger than that for unvoiced consonant parts. Therefore, we caluclate the likelihood of speaker models for a given noisy utterance using only a certain percentage frames having higher energy. This is our concept of skipping low SN frames. This is a special case of missing feature theory, that is, all features in a frame are missed. This approach is suitable for robust speaker identification.

4.3 Model Training Using Noisy Speech

Multicondition model training is effective for noisy environments as well as a clean environment. Therefore, we used a speaker model trained using utterances in noisy environments.

5. Experiments

5.1 Database and Speech Analysis

We used the NTT database [27], [30] and the JNAS (Japanese Newspaper Article Sentences) database [31] 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, 1991.6) in a sound proof room uttered at normal, fast and slow speaking style mode [30]. The content of the sentence utterance of NTT database is subset of the ATR 503 phonetic balance sentences. In this paper, only the sentences uttered at a normal speaking mode were used. Five same sentences for all speakers from one session (1990.8) were used to train speaker-specific GMMs. Our former study [26], [27] indicated that the speaker-specific GMM obtained slightly better performance than the speaker-adapted GMM in the case of 20 second training data. The speaker-specific GMM using small training data worked well because the number of mixtures of GMMs is relative small and the style of speech data is read speech. So we used the speaker-specific GMMs in this paper. Of course, the more the training data size is, the higher the speaker identification rate becomes. Five other sentences every from the other four sessions were used as test data. In other words, the test corpus consisted of $35 \times 5 \times 4 = 700$ trials for speaker identification. The average duration of the sentences is about 4 seconds. GMMs with 32 mixtures having diagonal covariance matrices were used as speaker models.

The scale of the NTT database is relatively small. The speaker experiment conducted on a large scale JNAS database [31] was also used to verify the robustness of our proposed phase-based method for speaker recognition in noisy environments. For the JNAS dababase, 270 speakers (135 males and 135 females) in the JNAS database were used for speaker identification. The content of the sentence utterance of JNAS database is Japanese newspaper article sentences. The reading text was made up of 150 sets which consisted of about 100 sentences each. Each speaker read one of 150 sets. All sentences were collected with headset microphone. 10 sentences (about 2 seconds[†]/sentence) were used for training speaker-specific GMMs, and 90 sentences (about 5.5 seconds[†] /sentence) were used for test. The test corpus consisted of $270 \times 90 = 24300$ trials for speaker identification. GMMs with 128 mixtures having diagonal covariance matrices were used as speaker models.

To obtain the noisy speech, we added stationary noise (in a computer room) and non-stationary noise (in an exhibition hall) to the utterance at the average SN ratios of 20 dB and 10 dB, respectively. The input speech was sampled at 16 kHz. 12 MFCCs for the NTT database[††]and 25 MFCCs (12 MFCCs and their first-order derivatives plus the first derivative of the power component) for the JNAS database were calculated at every 10 ms with a window of 25 ms [†††]. The spectrum with 128 components consisting of magnitude and phase was calculated every 5 ms with a window of 12.5 ms [††††]. FFT for 256 sampling points is performed, and we can get 128 components (128 real components and 128 imaginary components), in other words, 128 magnitude and 128 phase components. Literature [23] showed that the first 12 feature parameters, that is, from the $1st$ component (line spectrum: 8000/128 Hz) to $12th$ component (line spectrum: $8000/128 \times 12$ Hz) of the spectrum (frequency range: 60 Hz – 700 Hz) achieved the best identification performance of all the other sub-band frequency ranges. In this frequency range, the phase information has richer speaker characteristics than the power/magnitude spectrum [22]. In this paper, the phase information obtained from the lowest 12 components of the sub-band spectrum was used to evaluate the robustness of phase in noisy conditions. In our previous study [23], speaker identification performance of the modified normalized phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ obtained from Eq. (7) was significantly better than that of the original normalized phase $\{\tilde{\theta}\}$ obtained from Eq. (3) for NTT database of clean speech. On the other hand, as shown in Fig. 2 (b), the conventional phases without normalization extracted from two different windows of the same vowel from the same people were quite a bit different despite the fact that they should be very similar. So we think that the speaker recognition of the conventional phase without normalization would be very low. Therefore, only the modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ is used in this paper. The modified phase $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ means that the phase values $\{\tilde{\theta}\}$ are transformed to coordinates by Eq. (7), resulting in double the number of parameters compared with phase $\{\tilde{\theta}\}$. So the dimensions of the phase information corresponding to the lowest 12 components were 24.

5.2 Speaker Identification Results

5.2.1 Speaker Identification on NTT Database

We conducted the speaker identification experiment using phase information on the NTT database.

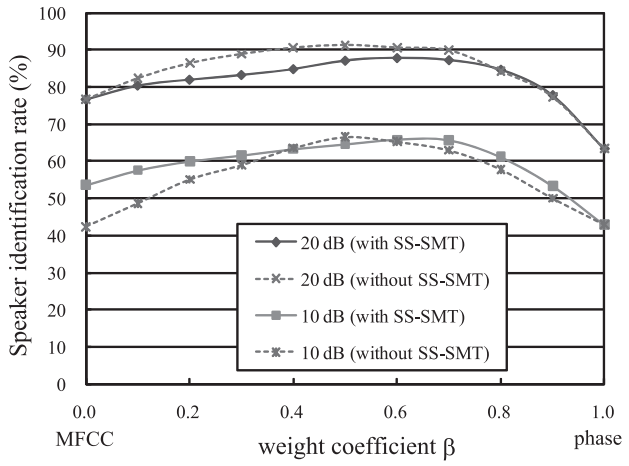
The speaker identification results by individual method and combination method by GMMs trained on clean speech are shown in Table 1 (a). For clean speech, although the performance of the phase-based method was worse than that of the MFCC-based method, the method has a basic speaker identification ability. As such, it is useful to use the phase information to identify a speaker. The combination of MFCCs and phase was significantly better than individual MFCC-based method. For noisy speech, the performance of phase information was similar to that of MFCC. The individual phase-based method even outperformed individual MFCC-

[†]Excluding about 2 seconds silence at beginning and ending of a sentence.

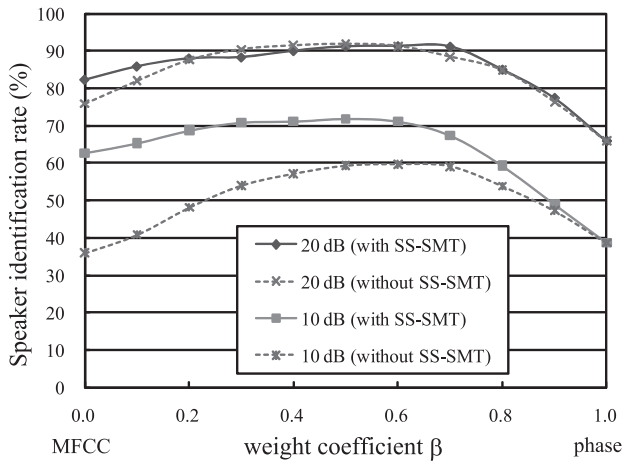
[††]In this paper, speaker identification experiment based on NTT database was used as a preliminary experiment. For the sake of convenience, 12 MFCCs were used. Of course, the better performance may be achieved if 25 MFCCs including Δ MFCCs will be used for NTT database (refer to Markov's paper [10]). However, the trend of our conclusion that the combination of the MFCC-based method and the phase information based method outperforms the MFCC-based method is the same.

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

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



(a) non-stationary noisy speech



(b) stationary noisy speech

Fig. 2 Speaker identification results of combination of MFCC with SS-SMT and Phase by GMMs trained on clean speech for NTT database.

Table 1 Speaker identification result by GMMs trained on clean speech (%).

test data	MFCC	Phase	comb	weight β
(a) NTT database				
clean speech	97.7	73.4	99.3	0.2
non-stationary 20 dB	76.7	63.4	91.3	0.5
non-stationary 10 dB	42.4	42.9	66.6	0.5
stationary 20 dB	75.9	65.9	91.9	0.5
stationary 10 dB	36.0	38.7	59.7	0.6
Average	**57.8**	**52.7**	**77.4**	
(b) JNAS database				
clean speech	98.5	88.7	98.9	
non-stationary 20 dB	51.0	60.8	81.6	0.7
non-stationary 10 dB	17.5	19.2	29.4	0.8
stationary 20 dB	36.4	72.0	72.0	1.0
stationary 10 dB	20.8	46.0	46.0	1.0
Average	**31.4**	**49.5**	**57.3**	

based method for low SNR (10 dB) speech. By combining phase and MFCC, the average identification error rate of stationary and non-stationary noisy speech was reduced to about 8.4% (100 − (91.3 + 91.6)/2) from 23.7% (100 − (76.7 + 75.9)/2) using only MFCC for 20 dB (relative error reduction rate of 64.6%) and to 36.8% (100−(66.6+59.7)/2) from 60.8% (100 − (42.4 + 36.0)/2) for 10 dB (relative error reduction rate of 39.5%), respectively. For noisy speech, optimal weight β was about 0.5. For the following experiments, almost the best combination results were obtained by setting weight β from 0.3 to 0.7. We did not add all weights β to Tables 2, 4 and 5 to keep the brevity of the paper.

Table 2 Speaker identification result by GMMs trained on noisy speech (%).

training data	test data	MFCC	Phase	comb
(a) NTT database				
non-stationary noise 20 dB	non-stationary 20	95.9	66.0	98.0
	non-stationary 10	89.3	47.3	93.7
	stationary 20	93.9	65.1	96.4
	stationary 10	81.1	43.4	88.7
	Average	**90.1**	**55.5**	**94.2**
non-stationary noise 10 dB	non-stationary 20	90.0	53.7	93.7
	non-stationary 10	93.1	51.7	95.3
	stationary 20	91.7	53.3	94.9
	stationary 10	87.7	49.7	91.3
	Average	90.6	52.1	93.8
stationary noise 20 dB	stationary 20	96.4	66.7	98.4
	stationary 10	88.6	46.0	93.6
	non-stationary 20	92.6	66.9	96.4
	non-stationary 10	89.6	45.6	92.0
	Average	**91.8**	**56.3**	**95.1**
stationary noise 10 dB	stationary 20	85.6	49.9	92.3
	stationary 10	94.1	45.1	96.0
	non-stationary 20	79.9	53.0	87.3
	non-stationary 10	88.3	46.6	92.3
	Average	87.0	48.7	92.0
(b) JNAS database				
non-stat noise 20 dB	non-stationary 20	97.6	73.0	97.8
	non-stationary 10	92.7	36.2	92.7
	stationary 20	21.2	55.1	57.2
	stationary 10	13.4	29.7	35.6
	Average	**56.2**	**48.5**	**70.8**
non-stat noise 10 dB	non-stationary 20	90.4	54.0	94.4
	non-stationary 10	95.7	40.0	95.7
	stationary 20	9.1	37.9	37.9
	stationary 10	5.8	23.5	25.4
	Average	50.3	38.9	63.4
stationary noise 20 dB	stationary 20	95.3	88.3	97.4
	stationary 10	86.0	66.8	90.9
	non-stationary 20	32.9	52.2	61.8
	non-stationary 10	16.0	18.8	27.2
	Average	**57.6**	**56.5**	**69.3**
stationary noise 10 dB	stationary 20	86.0	82.6	94.7
	stationary 10	92.2	81.7	95.4
	non-stationary 20	17.9	59.6	59.6
	non-stationary 10	12.1	25.0	25.0
	Average	52.1	62.2	68.7

Table 2 (a) shows the speaker identification results by GMMs trained on noisy speech. Speaker models trained on noisy speech were very effective especially for MFCC-based method and the combination method. For example, by using stationary noisy speech training data of 20 dB, the average identification rate was improved to 91.8% from 57.8% (relative error reduction rate of 80.6%) by clean speech training data for MFCC-based method. The combination result by noisy speech training data was also improved re-

Table 3 Speaker identification result of MFCC with SS-SMT by GMMs trained on clean speech (%).

α	20 dB		10 dB	
	non-stat	stat	non-stat	stat
(a) NTT database				
0.0	76.7	75.9	42.4	36.0
3.0	77.4	83.1	52.6	61.9
(b) JNAS database				
0.0	51.0	36.4	17.5	20.8
3.0	52.1	37.5	21.6	18.3

Table 4 Speaker identification results by deleting low energy frames by GMMs trained on clean speech (%).

test data	MFCC	Phase	comb
(a) NTT database			
non-stat 20 dB-10%	77.4	68.0	89.3
non-stat 10 dB-30%	56.1	54.3	78.0
stationary 20 dB-10%	84.0	69.9	94.9
stationary 10 dB-30%	49.3	54.9	78.1
Average	**66.7**	**61.8**	**85.1**
(b) JNAS database			
non-stat 20 dB-40%	81.3	73.3	94.0
non-stat 10 dB-40%	42.3	38.8	61.6
stationary 20 dB-40%	61.5	76.8	89.0
stationary 10 dB-40%	35.0	62.2	68.1
Average	**55.0**	**62.8**	**78.2**

markably compared to that by clean speech training data even when noise nature and SNR of the test data were mismatched to those of the training data.

Table 3 (a) summarizes the identification rates of NTT database based on MFCC using Spectrum Subtraction with SMoothing of Time direction (SS-SMT). The improvement was significant especially for low SNR and stationary noisy speech. For example, for the case of 10 dB, the rates were improved to 52.6% from 42.4% (relative error reduction of 17.7%) for non-stationary speech data, and to 61.9% from 36.0% (relative error reduction of 40.5%) for stationary speech data, respectively. The combination results of MFCC with SS-SMT and the phase information of NTT database are shown in Fig. 2. The combination results were remarkably better than that of individual MFCC-based method. By integrating MFCC with phase, the result with SS-SMT processing outperformed the result without SS-SMT processing for stationary noisy speech of 10 dB. However, the improvement was not achieved in other cases.

Table 4 (a) shows speaker identification results of NTT database by deleting (skipping) frames having low energy for clean speech training data. The notation of 20 dB-10% denotes that 10% frames having the lowest energy for 20 dB noisy speech is deleted. The rate of deleted frames was determined empirically. The speaker identification performance by deleting (skipping) low energy frames achieved a significant improvement than that using all frames for MFCC, the phase information and the combination. The combination result was also significantly better than that of MFCC-based method. By comparing Table 1 with Table 4 (a), we can see that the function of deletion for unreliable frames (frames having low energy) improved the average identification rate of combination result to 85.1% from 77.4%.

Table 5 Speaker identification results by deleting low energy frames by GMMs trained on stationary/non-stationary noisy speech of 20 dB (%).

training data	test data	MFCC	Phase	comb
(a) NTT database				
non-stat noise 20 dB	non-stat 20 dB-10%	93.7	61.7	96.9
	non-stat 10 dB-30%	89.3	49.6	91.3
	stat 20 dB-10%	80.6	62.3	89.0
	stat 10 dB-30%	81.6	45.7	84.9
	Average	86.3	54.8	90.5
stationary noise 20 dB	stationary 20 dB-10%	96.4	61.7	98.7
	stationary 10 dB-20%	88.3	47.0	93.4
	non-stat 20 dB-10%	94.9	61.3	98.0
	non-stat 10 dB-20%	89.9	48.4	93.7
	Average	**92.4**	**54.6**	**96.0**
(b) JNAS database				
non-stat noise 20 dB	non-stat 20 dB-40%	97.9	75.8	98.3
	non-stat 10 dB-40%	95.3	54.2	95.3
	stationary 20 dB-40%	45.7	72.4	83.9
	stationary 10 dB-40%	26.6	55.4	63.0
	Average	**66.4**	**64.5**	**85.2**
stationary noise 20 dB	stationary 20 dB-40%	96.2	79.9	97.9
	stationary 10 dB-40%	90.5	71.3	95.1
	non-stat 20 dB-40%	43.4	70.9	74.9
	non-stat 10 dB-40%	29.7	36.3	47.6
	Average	65.0	64.6	78.9

Finally, Table 5 (a) shows the speaker identification results of NTT database by deleting low energy frames for noisy training data of 20 dB. Using GMMs trained on stationary noisy speech of 20 dB, we find that the deletion of unreliable frames improved the identification rate from 96.4% (see Table 2 (a)) to 98.0% for 20 dB non-stationary noisy speech and from 92.0% (see Table 2 (a)) to 93.7% for 10 dB non-stationary noisy speech. The rates based on only MFCC were 76.7% and 42.4% by using clean speech training data, respectively (see Table 1), that is, the error reduction rates were 91.4% and 89.1%, repectively. The combination also achieved a relative error reduction rate of 47.4% compared to the individual MFCC-based best result (92.4%). The results based on GMMs trained on non-stationary noisy speech of 20 dB were worse than those based on GMMs trained on stationary noisy speech of 20 dB. For GMMs trained on non-stationary noisy speech of 20 dB, the results by deleting low energy frames were worse than those using all frames. We could not find the reason of degradation. It remains for our future work.

5.2.2 Speaker Identification on Large Sacle JNAS Database

The speaker identification experiment using phase information on a large scale JNAS database [31] was also conducted in this section.

The speaker identification results on the JNAS database using clean speech are shown in Table 1 (b). The phase information was also very effective for the large scale JNAS database especially for stationary noisy speech. The indi-

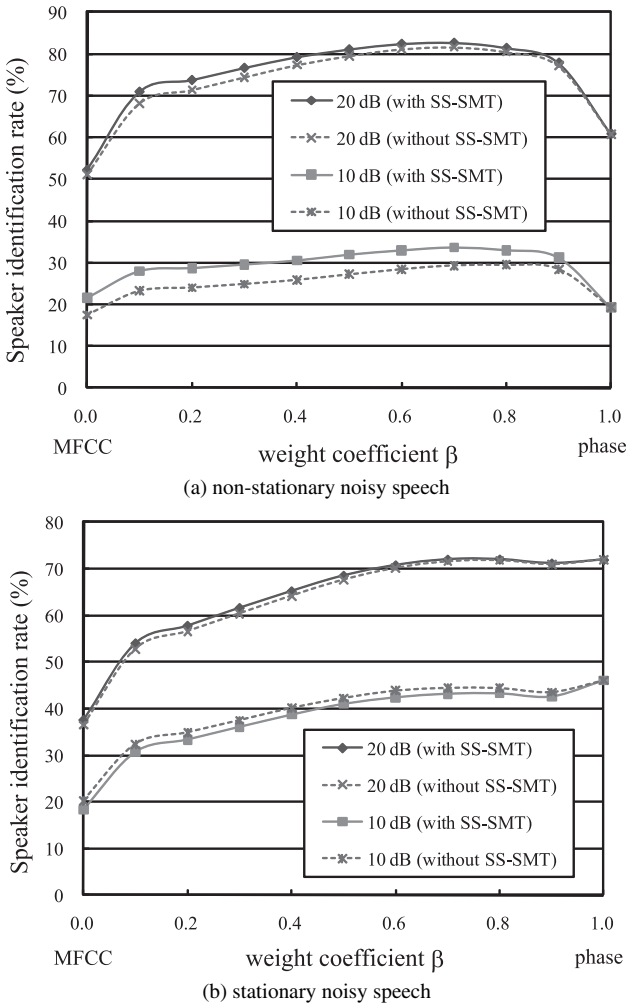


Fig. 3 Speaker identification results of combination of MFCC with SS-SMT and Phase by GMMs trained on clean speech for JNAS database.

vidual results of the phase information outperformed that of MFCCs under all noisy conditions, and an average relative error reduction of 26.4% (identification rate of MFCC: 31.4% → that of phase:49.5%) was achieved. For stationary noisy speech, the recogniton rate of the phase information was about 2 times as that of MFCCs. By integrating MFCCs with the phase information, an average relative error reduction of 37.8% (31.4% → 57.3%) was achieved.

Table 2 (b) shows the speaker identification results on the JNAS database using noisy training data. Using speaker models trained on noisy speech, the identification performances of MFCC-based method and the combination method were improved significantly. For example, by using non-stationary noisy speech training data of 20 dB, the result of combination of MFCC and phase (70.8%) achieved a relative error reduction rate of 33.3% over that of MFCC-based method (56.2%).

The resultss of individual MFCC with SS-SMT and the combination results of MFCC with SS-SMT and the phase information for JNAS database are shown in Table 3 (b) and Fig. 3, respectively. The combination results were remarkably better than that of individual MFCC-based method. For individual MFCC or combination of MFCC and phase, the results with SS-SMT processing worked worse than the results without SS-SMT processing for stationary noisy speech of 10 dB. On the other hand, MFCC with SS-SMT outperformed MFCC without SS-SMT in other cases. The improvement or the degradation between MFCC with SS-SMT and MFCC without SS-SMT was not very remarkbale.

Table 4 (b) shows speaker identification results of JNAS database by deleting (skipping) low energy frames (unreliable frames) for clean speech training data. The speaker identification performance by deleting low energy (low SNR) frames achieved a significant improvement than that using all frames. The result of the phase information was significantly better than that of MFCC especially for stationary noisy speech. The combination result (78.2%) achieved a relative error reduction of more than 50.0% over the result based on MFCCs (55.0%).

Table 5 (b) shows the speaker identification results of JNAS database by deleting low energy (low SNR) frames for stationary/non-stationary noisy training data of 20 dB. By comparing with Table 2 (b), we find that the deletion of unreliable frames improved the identification rate from 56.2% to 66.4% for MFCCs, from 48.5% to 64.5% for the phase information and from 70.8% to 85.2% for the combination result by using 20 dB non-stationary noisy speech training data. The relative error reduction rate was 78.4% over MFCC-based method using clean speech training data (31.4%: see Table 1 (b) → 85.2%) and 56.0% over the individual MFCC-based best result (66.4% → 85.2%).

6. Conclusion

In this paper, we investigated the robustness of the proposed phase information for speaker identification in noisy environments. A method skipping low energy frames or unreliable frames, speaker model trained on noisy speech and spectral subtraction were used to analyze the effect of the phase information and MFCCs under match and mismatch training conditions. To verify the robustness of our proposed phase-based method, the small scale NTT database and the large scale JNAS database added with stationary/non-stationary noise were performed. The experimental results were summarized in Fig. 4. We have following conclusions:

- The degradation of the phase information was smaller than that of MFCCs in noisy conditions especially for mismatch training models. The individual result of the phase information was even better than that of MFCCs by clean speech training models in many cases.
- The combination of MFCC and phase information improved the speaker identification performance remarkably than MFCC-based method. By combining MFCCs and the phase information, the relative error reduction rate was about 50.0% over the individual MFCC-based best result (see Table 5).
- By deleting unreliable frames (frames having low en-

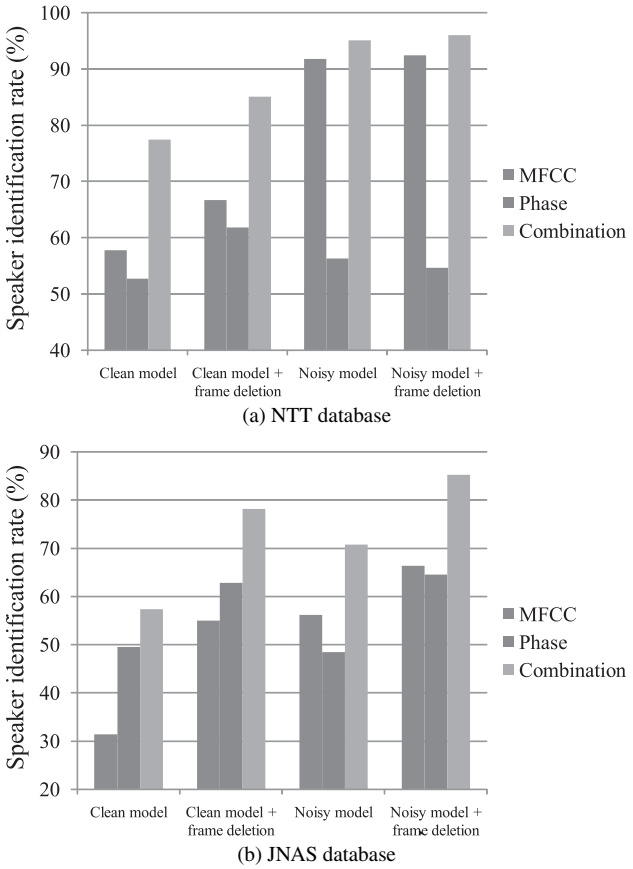


Fig. 4 Summary of speaker identificaton results of NTT database and JNAS database in stationary/nonstationary noisy environments (10/20 dB). Noisy model: speaker models trained on stationary noisy speech of 20 dB for NTT database and non-stationary noisy speech of 20 dB for JNAS database. Frame deletion: Deleting unreliable frames (frames having low energy/SN).

ergy), the speaker identification performance was improved significantly.

- The proposed methods were effective for both of the NTT database (small-scale, but different recording sessions) and JNAS database (large-scale).

That is to say, phase information is very robust and effective for speaker recognition in noisy conditions and it is an effective complementary feature of MFCCs. MFCC feature extracted from the power spectrum of noisy speech was greatly degraded by adding the power spectrum of clean speech with the power spectrum of noise. We assume that the phase of noise is random and the average of the phase of noise is near to zero. Thus, the phase was relatively more robust than the power spectrum based feature such as MFCC in noisy conditions. We also showed the effectiveness of the combination method of MFCC-based GMM and MFCC-based HMM [26], [27]. In future, we will investigate the combination methd of MFCC-based GMM, MFCC-based HMM and phase-based GMM.

Acknowledgments

This work was partially supported by a research grant from the Research Foundation for the Electrotechnology of Chubu (REFEC).

References

[1] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," Proc. ICASSP, pp.121–124, 1998.
[2] J. Ming, T.J. Hazen, J.R. Glass, and D.A. Reynalds, "Robust speaker recognition in noisy conditions," IEEE Trans. Audio, Speech Language Process, vol.15, no.5, pp.1711–1723, 2007.
[3] D. Pullella, M. Kuhne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," Proc. ICASSP, pp.4833–4836, 2008.
[4] T. Kamada, N. Minematsu, T. Osanai, H. Makinae, and M. Tanimoto, "Speaker verification in realistic noisy environment in forensic science," IEICE Trans. Inf. & Syst., vol.E91-D, no.3, pp.558–566, March 2008.
[5] T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using HMM composition in noisy environments," Comput. Speech Lang., vol.10, no.2, pp.107–116, 1996.
[6] L.P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," Proc. ICASSP, vol.1, pp.457–460, 2001.
[7] L. Gonzalez-Rodriguez and J. Ortega-Garcia, "Robust speaker recognition through acoustic array processing and spectral normalization," Proc. ICASSP, pp.1103–1106, 1997.
[8] I. McCowan, J. Pelecanos, and S. Scridha, "Robust speaker recognition using microphone arrays," Proc. A Speaker Odyssey–The Speaker Recognition Workshop, pp.101–106, 2001.
[9] T. Asami, K. Iwano, and S. Furui, "Evaluation of a noise-robust multi-stream speaker verification method using F0 information," IEICE Trans. Inf. & Syst., vol.E91-D, no.3, pp.549–557, March 2008.
[10] K.P. Markov and S. Nakagawa, "Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition," J. Acoust. Soc. Jpn. (E), vol.20, no.4, pp.281–291, 1999.
[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, 2006.
[12] N. Zheng, T. Lee, and P.C. Ching, "Integration of complementary acoustic features for speaker recognition," IEEE Signal Process. Lett., vol.14, no.3, pp.181–184, 2007.
[13] W.N. Chan, N. Zheng, and T. Lee, "Discrimination power of vocal source and vocal tract related features for speaker segmentation," IEEE Trans. Audio, Speech Language Process., vol.15, no.6, pp.1884–1892, 2007.
[14] K.K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," Proc. Eurospeech-2003, pp.2117–2120, 2003.
[15] G. Shi, et al., "On the importance of phase in human speech recognition," IEEE Trans. Audio, Speech Language Process., vol.14, no.5, pp.1867–1874, 2006.
[16] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," Proc. ICASSP, vol.1, pp.133–136, 2001.
[17] P. Aarabi, et al., Phase-based speech processing, World Scientific, 2005.
[18] R.M. Hegde, H.A. Murthy, and G.V.R. Rao, "Application of the modified group delay function to speaker identification and discrimination," Proc. ICASSP, pp.517–520, 2004.

[19] R. Padmanabhan, S. Parthasarathi, and H. Murthy, "Robustness of phase based features for speaker recognition," Proc. INTERSPEECH, pp.2355–2358, 2009.
[20] J. Kua, J. Epps, E. Ambikairajah, and E. Choi, "LS regularization of group delay features for speaker recognition," Proc. INTERSPEECH, pp.2887–2890, 2009.
[21] http://www.nist.gov/speech/tests/sre/
[22] S. Nakagawa, K. Asakawa, and L. Wang, "Speaker recognition by combining MFCC and phase information," Proc. Interspeech, pp.2005–2008, 2007.
[23] L. Wang, S. Ohtsuka, and S. Nakagawa, "High improvement of speaker identification and verification by combining MFCC and phase information," Proc. ICASSP, pp.4529–4532, 2009.
[24] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Commun., vol.17, no.1-2, pp.91–108, 1995.
[25] D.A. Reynolds, T.F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Digit. Signal Process., vol.10, no.1-3, pp.19–41, 2000.
[26] S. Nakagawa, W. Zhang, and M. Takahashi, "Text-independent speaker recognition by combining speaker specific GMM with speaker adapted syllable-based HMM," Proc. ICASSP, vol.I, pp.81–84, 2004.
[27] 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. Inf. & Syst., vol.E89-D, no.3, pp.1058–1064, March 2006.
[28] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, no.2, pp.113–120, April 1979.
[29] N. Kitaoka and S. Nakagawa, "Evaluation of spectral subtraction with smoothing of time direction on the AURORA2 task," Proc. ICSLP2002, pp.465–468, 2002.
[30] T. Matusi and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," Proc. ICASSP'93, vol.II, pp.391–394, 1993.
[31] 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.

Longbiao Wang received his B.E. degree from Fuzhou University, China, in 2000 and an M.E. and Dr. Eng. degree from Toyohashi University of Technology, 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 at Shizuoka University, Japan. His research interests include robust speech recognition, speaker recognition and source localization. He is a member of IEEE, Institute of Electronics and Acoustical Society of Japan (ASJ).

Kazue Minami received her B.E. degree from Toyohashi University of Technology in 2009. She is now a master student at Toyohashi University of Technology. Her research interests include speaker recognition.

Kazumasa Yamamoto received his B.E., M.E. and Dr. of Eng. degrees in Information and Computer Sciences from Toyohashi University of Technology, Toyohashi, Japan, in 1995, 1997, and 2000, respectively. From 2000 to 2007, he was a Research Associate in the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shinshu University, Nagano, Japan. Since 2007, he has been an Assistant Professor in the Department of Information and computer Sciences, Toyohashi University of Technology. His current research interests include speech recognition and privacy protection for speech signals. He is a member of ASJ, and Information Processing Society of Japan (IPSJ).

Seiichi Nakagawa received a 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. He moved to Toyohashi University of Technology in 1980. 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. From 1985 to 1986, he was a Visiting Scientist in the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, USA. He received the 1997/2001 Paper Award from the IEICE and the 1988 JC Bose Memorial Award from the Institution of Electro. Telecomm. Engrs. His major interests in research include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence. He is a fellow of IPSJ.