Speaker Identification Using Pseudo Pitch Synchronized Phase Information in Noisy Environments

Yuta Kawakami*, Longbiao Wang*, and Seiichi Nakagawa†
* Department of Electrical Engineering, Nagaoka University of Technology, Japan
E-mail: s123118@stn.nagaokaut.ac.jp, wang@vos.nagaokaut.ac.jp
† Department of Computer Science and Engineering, Toyohashi University of Technology, Japan

***Abstract*—In conventional speaker identification methods based on mel-frequency cepstral coefficients (MFCCs), phase information is ignored. Recent studies have shown that phase information contains speaker dependent characteristics, and, pitch synchronous phase information is more suitable for speaker identification. In this paper, we verify the effectiveness of pitch synchronous phase information for speaker identification in noisy environments. Experiments were conducted using the JNAS (Japanese Newspaper Article Sentence) database. The pseudo pitch synchronized phase information based method achieved a relative speaker identification error reduction rate of 15.5% compared to the conventional phase information (that is pitch non-synchronized phase). By cutting frames with low power and combining phase information with MFCC, a furthermore improvement was obtained.**

I. Introduction

In conventional speaker identification methods based on mel-frequency cepstral coefficients (MFCCs), only the magnitude of the Fourier Transform in time-domain speech frames has been used. This means that the phase component is ignored. MFCCs capture not only speaker-specific vocal tract information, but also some vocal source characteristics. However, speaker characteristics in the voice source are not captured completely by the MFCC. Therefore, feature parameters extracted from excitation source characteristics are also useful for speaker identification [1]-[6]. Almost all of the existing methods are based on Linear Predictive Coding (LPC) analysis. Markov and Nakagawa proposed a Gaussian Mixture Model (GMM) based text-independent speaker identification system that integrates pitch and the LPC residual with the LPC-derived cepstral coefficients [2]. Their experimental results show that using pitch information is the most effective when the correlation between pitch and the cepstral coefficients is taken into consideration. An automatic technique for estimating and modeling the glottal flow derivative source waveform of speech and applying the model parameters to speaker identification was proposed in [3]. The complementary nature of speaker-specific information in the residual phase compared with the information in conventional MFCCs was demonstrated in [4]. The residual phase was derived from speech signals by linear prediction analysis. Zheng et al. proposed a speaker verification system using complementary acoustic features derived from vocal source excitation and the vocal-tract system [5]. A new feature set, called the wavelet octave coefficients of residues (WOCOR), was proposed to capture the spectro-temporal source excitation characteristics embedded in the linear predictive residual signal [5]. Recently, many speaker recognition studies using group delay based phase information have been proposed [7], [8]. Wang et al. proposed phase-related features for speaker recognition [9]. This type of phase information considers all frequency ranges. We think that phase information is valid for speaker identification, since it captures the features of the source wave.

Previously, we proposed a speaker identification system using a combination of MFCCs and phase information [1], [10], [11], directly extracted from the limited bandwidth of the Fourier transform of the speech wave. We also showed that the phase information is effective for speaker identification in clean and noisy environments [1], [10], [11], [12]. However, problems occurred in extracting the phase information because of the influence of the windowing position. Therefore, we propose a new method to extract pitch synchronous phase information and skip frames with low power [15].

In this paper, we show that the pseudo pitch synchronous phase information is also affective for noisy speech. The rest of this paper is organized as follows. Section 2 presents the phase information extraction method, while Section 3 discusses combining the phase and MFCC methods. The experimental setup and results are reported in Section 4, and Section 5 presents our conclusions.

II. Phase information extraction

A. Formulas [1], [12]

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

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

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

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

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

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

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

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

and the phase information is normalized as follows:

$$\tilde{\theta}(\omega',t) = \theta(\omega',t)+\frac{\omega'}{\omega}(\frac{\pi}{4}-\theta(\omega,t)) \quad (6)$$

In the experiments described in this paper, the basis frequency ω is set to $2\pi \times 1000 Hz$. In a previous study, to reduce the number of feature parameters, we used phase information in a sub-band frequency range only. However, a problem arose with this method when comparing two phase values. For example, for two values $\pi - \tilde{\theta}_1$ 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 each other. Therefore, we modified the phase into coordinates on a unit circle [12], that is,

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

B. Improvement of phase information extraction

Using the relative phase extraction method that normalizes the phase variation by cutting positions, we can reduce the phase variation. However, the normalization of phase variation is still inadequate. For example, for a 1000 Hz periodic wave (16 samples per cycle for a 16 kHz sampling frequency), if one sample point shifts in the cutting position, the phase shifts only $\frac{2\pi}{16}$, while for a 500 Hz periodic wave, the phase shifts only $\frac{2\pi}{32}$ with this single sample cutting shift. On the other hand, if the 17 sample points shift, their phases will shift by $\frac{17\cdot 2\pi}{16}(mod 2\pi) = \frac{2\pi}{16}$ and $\frac{34\pi}{32}$, respectively, for the two periodic waves. Therefore, the values of the relative phase information for different cutting positions are very different from those of the original cutting position. We have addressed such variations using a statistical distribution model of GMM [1], [10], [12].

If we could split the utterance by each pitch cycle, changes in the phase information would be further obviated. Thus, we propose a new extraction method that synchronizes the splitting section with a pseudo pitch cycle [15].

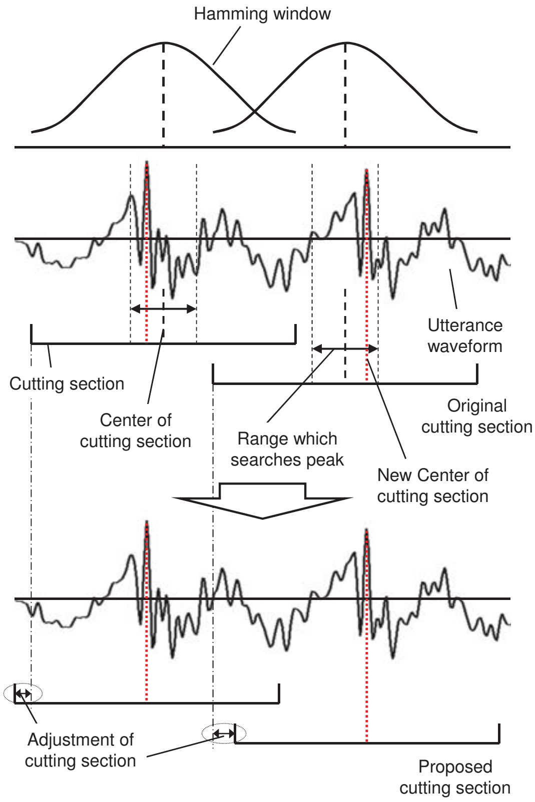
Hamming window
Utterance waveform
Cutting section
Center of cutting section
Range which searches peak
Original cutting section
New Center of cutting section
Adjustment of cutting section
Proposed cutting section

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

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

C. Cutting low power frames [12]

In noisy environments, speaker recognition performance is degraded significantly. To address this problem, we cut 40% of frames that have low power in the speech. This means we use 60% of frames in the speech. This method obtain the power from all frames in the speech, and sort the value. Then, frames lower 40% are unused. In the previous study[12], this method was effective for improving speaker identification performance using MFCC and the conventional pitch non-synchronized phase information. We apply this method to only test data, not to the training data.

III. Combination method and decision method

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

L^n_{comb} given by

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

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

IV. Experiments

A. Database and speech analysis

We used the JNAS (Japanese Newspaper Article Sentence) database in the experiments [13]. The JNAS corpus consists of the recordings of 270 speakers (135 males and 135 females). To train the models, 10 clean sentences were used for all speakers. About ninety other sentences were used as test data. To obtain the noisy speech, we added stationary noise (in a computer room) and non-stationary noise (in an exhibition hall) from JEIDA Noise Database [14] to the test speech at the average SN (Signal-to-Noise) ratios of 20 dB and 10 dB. In total, the test corpus consisted of about 24,000 (90×270) trials for each condition. The average duration of the sentences is approximately 5.5 seconds.

The input speech was sampled at 16 kHz. A total of 25 dimensions (12 MFCCs, 12 ΔMFCCs and Δpower) were calculated every 10 ms with a window length of 25 ms. The spectrum with 128 components consisting of magnitude and phase was calculated by DFT for every 256 samples. Phase information was calculated every 5 ms with a window length of 12.5 ms. For phase information, we used the first 12 phase components (24 feature parameters in total), that is, from the first to the 12th component of the phase spectrum (frequency range: 62.5 Hz - 750 Hz), which achieved the best identification performance among all the other sub-band frequency ranges [1]. These analysis conditions are shown in Table I briefly.

TABLE I
Analysis conditions for MFCC and Phase information

	MFCC	Phase
Frame length	25ms	12.5ms
Frame shift	10ms	5ms
FFT size	512 samples (400 datas plus 112 zeros)	256 samples (200 datas plus 56 zeros)
Dimensions	25 (12MFCCs, 12MFCCs, and Δpower)	24 (sinθ and cosθ of the first 12 components of the phase spectrum)

B. Speaker identification results

We conducted a speaker identification experiment using phase information on the JNAS database. GMMs with 128 mixtures were used as speaker models. The new phase extraction method searches for the peak amplitude point in the range ±0 ms, ±2.5 ms in the center of the next window. The speaker identification results obtained from the individual methods and the combination methods are shown in Table II and Fig.2. "±0 ms" corresponds to the conventional extraction method. For example, in the conventional method, recognition rate is 68.6% for stationary noisy speech of 20 dB. On the other hand, using the newly proposed extraction technique (Phase ±2.5ms), the recognition rate is improved to 81.0%. On average, comparing with the conventional phase information, the speaker identification rate improved from 47.2% to 55.4% (that is, an average error reduction rate is 15.5%). This means that the window could catch the pitch accurately by searching peaks in many frames, and more effective phase informations were extracted even it was in the noisy environments.

The speaker identification results obtaind from the cutting sections method are shown in Table III and Fig. 3. Comparing with Table II, the recognition rates were improved in the all conditions. The reason was that unreliable likelihood of frames with low power was deleted. The proposed pseudo pitch synhronized phase information outperformed than the conventional phase information in almost all cases. However, the proposed phase information was little worse (from 24.5% to 23.4%) than the conventional phase information for non-stationary noisy speech of 10 dB. The reason might be that the positions of some sudden large noise were mistaken as center of the window. It is interesting that MFCC is more robust for non-stationary noise than stationary noise, but that phase information is the opposite. Therefore, these features are comprementary each other.

TABLE II
Speaker identification results in noisy environments using all frames in the speech ("stat" means stationary noise in a computer room, "non-stat" means non-stationary noise in a exhibision hall, decimals in ( ) are α in equation (8))

	noise condition		average
	10 dB stat / non-stat	20 dB stat / non-stat	
MFCC	20.2 / 18.2	36.4 / 52.4	31.8
Phase ±0ms	43.1 / 19.7	68.6 / 57.5	47.2
Phase ±2.5ms	56.2 / 17.0	81.0 / 67.4	55.4
combination MFCC and Phase ±0ms	43.1 / 29.1 (1.0) (0.7)	69.7 / 78.5 (0.8) (0.6)	55.1
combination MFCC and Phase ±2.5ms	56.2 / 27.1 (1.0) (0.6)	81.0 / 82.1 (1.0) (0.6)	61.6

TABLE III
Speaker identification results in noisy environments using 40% cut frames in the speech

	noise condition		average
	10 dB stat / non-stat	20 dB stat / non-stat	
MFCC	35.0 / 42.3	61.5 / 81.3	55.0
Phase ±0ms	52.4 / 24.5	76.4 / 67.7	55.3
Phase ±2.5ms	65.4 / 23.4	86.4 / 78.0	63.3
combination MFCC and Phase ±0ms	60.3 / 53.1 (0.6) (0.5)	87.9 / 92.1 (0.7) (0.6)	73.4
combination MFCC and Phase ±2.5ms	67.0 / 52.3 (0.7) (0.4)	90.1 / 94.2 (0.7) (0.5)	75.9

The speaker identification results obtained from the combination method are summarized in Fig. 4. For example, when the MFCC-based method is compared with the combination

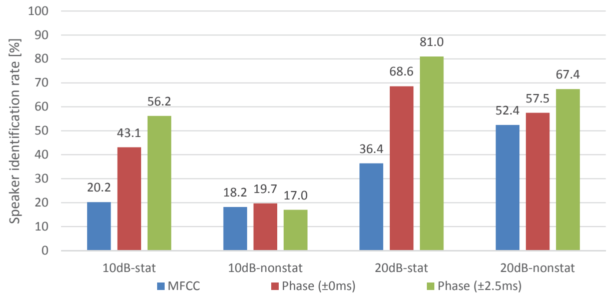

Fig. 2. *Speaker identification results using MFCC and phase (use all frames in the speech).*

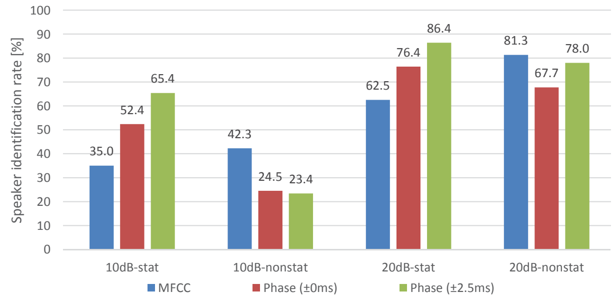

Fig. 3. *Speaker identification results using MFCC and phase (cut 40% of frames in the speech).*

method for stationary noisy speech of 10 dB, the rate was improved from 20.2% to 56.2% (a relative error reduction of 45.1%) for the conventional phase extraction method, and from 43.1% to 56.2% (a relative error reduction rate of 23.0%) for the new phase extraction method using all frames in the speech. On average, the rate was improved from 55.1% to 61.6% when using all frames of the speech (see the Table II). Moreover, by cutting 40% of frames with low power in the speech, the rate was improved to 75.9% on average (see the Table III).

This result suggests that the proposed pseudo pitch synchronized phase information extraction method is more effective than the conventional extraction method in noisy conditions.

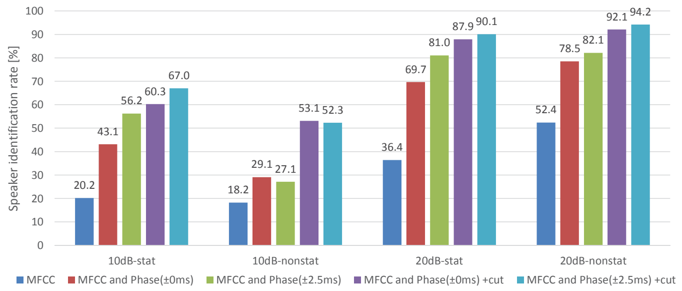

Fig. 4. *Speaker identification results using combination of MFCC and phase.*

V. Conclusions and Future Works

In this paper, we verified the effectiveness of a pseudo-pitch synchronous phase information for speaker identification in noisy environments. Using the proposed method, on average, the speaker recognition rate was improved from 47.2% using the conventional phase information to 55.4% by individual method. Moreover, the recognition rate using the MFCC was improved remarkably when combined with the proposed phase information (from 31.8% to 61.6%). And by cutting low power sections in the speech, the rate was improved to 75.9%. These results confirm that the proposed phase information is useful for speaker identification in noisy environments.

But this method has weakness at large and non-stationary noise conditions. In our preliminary experiment, the performance of pseudo pitch synchronized phase information of noisy speech using the estimated peak position of clean speech (that is, an ideal condition) was better than that using the estimated peak position of noisy speech. This means that one of the weakness is the mistaking of estimation of the peak position. In the proposed method, the center of the frame is decided only depending on the peak amplitude of the signal. We will try to find a robust method to detect the peak position of voice wave. To find the true peak of voice wave, the moving average method is very simple but might be useful because it can reduce sudden peak of noise.

References

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

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

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

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

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

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

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

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

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

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

[11] S. Nakagawa, L. Wang and S. Ohtsuka, “Speaker identification and verification by combining MFCC and phase information”, IEEE Trans. on Audio, Speech, and Language processing, Vol. 20, No. 4, pp.1085-1095 (2012).

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

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

[14] http://research.nii.ac.jp/src/en/JEIDA-NOISE.html

[15] K. Shimada, K. Yamamoto and S. Nakagawa, “Speaker identification using pseudo pitch synchronized phase information in voiced sound”, Proc. APSIPA, pp, 1-6 (2010).