On the Use of Phase Information-based Joint Factor Analysis for Speaker Verification under Channel Mismatch Condition

Ikuya Hirano*, Longbiao Wang†, Atsuhiko Kai* and Seiichi Nakagawa‡
* Shizuoka University, Japan
E-mail: hirano@spa.sys.eng.shizuoka.ac.jp, kai@sys.eng.shizuoka.ac.jp
† Nagaoka University of Technology, Japan
E-mail: wang@vos.nagaokaut.ac.jp
‡ Toyohashi University of Technology, Japan
E-mail: nakagawa@slp.cs.tut.ac.jp

***Abstract*—Recent studies have shown that phase information contains speaker characteristics. A new extraction method to extract pitch synchronous phase information has been proposed and shown that it was very effective under channel matched condition. However, phase changes between different channels. Therefore, the speaker recognition performance is drastically degraded under channel mismatch condition. On the other hand, joint factor analysis (JFA) is an approach that is robust for channel variability. In this paper, we propose phase information-based JFA for speaker verification under channel mismatch condition. Speaker verification experiments were performed using the NIST 2003 SRE database. Phase information-based JFA achieved a relative equal error rate reduction of 20.9% for male and 17.4% for female compared to the traditional system based on Gaussian mixture model and Universal background model (GMM-UBM) that influenced by channel variability. Furthermore, by combining phase information-based method with the MFCC-based method, we obtained the better result than that of the only MFCC-based method.**

I. Introduction

In conventional speaker verification 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 has been ignored. Importance of phase in human speech recognition has been reported in [1], [2]. Several studies have invested great effort in modeling and incorporating the phase into the speaker recognition process [3]. The complementary nature of speaker-specific information in the residual phase compared with the information in conventional MFCCs was demonstrated in [3]. 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 [4], [5].

Previously, Wang et al. proposed a speaker verification system using a combination of MFCCs and phase information [6], [7] directly extracted from the limited bandwidth of the Fourier transform of the speech wave. However, problems occurred in extracting the phase information because of the influence of the windowing position. Shimada et al. proposed a new method to extract pitch synchronous phase information [8]. The experimental results showed that the phase information was effective for speaker recognition under channel matched condition [6], [7], [8].

However, phase drastically changes between different channels. In [9], the experimental results indicated that the speaker recognition performance based on phase information was drastically degraded under channel mismatch and channel distortion conditions. To mitigate the influence of channel mismatch for phase information, joint factor analysis (JFA) [10] instead of traditional GMM-UBM based on Gaussian mixture model (GMM) and Universal background model (UBM) is used in this study. Recently, the JFA approach has become the active field for speaker verification. This modeling proposes powerful tools for addressing the problem of speaker and channel variability in GMM framework. Therefore, it is considered that the degradation of speaker verification performance using phase information under channel mismatch condition would be mitigated partly. Furthermore, a combination of the phase information-based JFA and MFCC-based JFA is also studied in this paper.

II. Phase information extraction

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

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

However, the phase $\theta(\omega, t)$ changes according to the frame position in the input speech. To overcome the influence of the phase response with respect to frame position, phases with the anchoring radian frequency ω_b for all frames are converted to a constant, and the phase with the other frequency is estimated relative to this. In the experiments discussed in this paper, the anchoring radian frequency ω_b is set to $2\pi \times 1000$ Hz. Actually, this constant phase value of the anchoring radian frequency does not affect the speaker recognition result. Without loss of generality, setting the phase with the anchoring radian frequency $\theta(\omega_b, t)$ to 0, we have

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

whereas for the other frequency $\omega = 2\pi f$, the spectrum on frequency ω is normalized as

$$S'(\omega, t) = \sqrt{X^2(\omega, t) + Y^2(\omega, t)} \times e^{j\theta(\omega, t)} \times e^{j\frac{\omega}{\omega_b}(-\theta(\omega_b, t))}. \quad (3)$$

Then, the phase information is normalized as

$$\tilde{\theta}(\omega, t) = \theta(\omega, t) + \frac{\omega}{\omega_b}(-\theta(\omega_b, t)). \quad (4)$$

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 mapped the phase into coordinates on a unit circle [6], [7], that is,

$$\tilde{\theta} \rightarrow \{cos\tilde{\theta}, sin\tilde{\theta}\}. \quad (5)$$

Using the relative phase extraction method that normalizes the phase variation with respect to frame positions, the phase variation can be reduced. 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 (frame 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 [6], [7].

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

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

In this paper, however, we don't discuss the comparison with traditional phase information and pseudo-pitch synchronous phase information because the effectiveness of pseudo-pitch synchronous phase information has already been shown in [8].

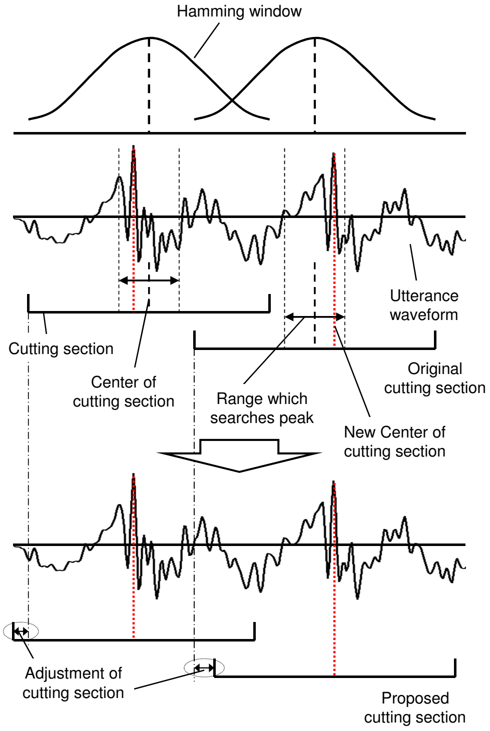
Hamming window
Utterance waveform
Cutting section
Center of cutting section
Range which searches peak
Original cutting section
New Center of cutting section
Adjustment of cutting section
Proposed cutting section

Fig. 1. How to synchronize the splitting section

III. Joint factor analysis

Joint factor analysis is an effective model for speaker verification under channel mismatch conditon. In this model, each speaker is represented by the means, covariance, and weights of a mixture of multivariate diagonal-covariance Gaussian densities defined in some continuous feature space of dimensions. The GMM for a target speaker is derived by adapting the universal background model (UBM) mean parameters. The basic assumption in JFA is shown as (6).

$$M = s + c, \quad (6)$$

where M is a speaker- and channel-dependent supervector, and s and c are speaker and channel supervectors, respectively.

The first term in the right-hand side of (6) is modeled by supposing that if s is the speaker supervector for a rondomly chosen speaker, then

$$s = m + Dz + Vy, \quad (7)$$

where m is the speaker- and channel-independent supervector (UBM), D is a diagonal matrix, V is a rectangular matrix of low rank, and y and z are independent random vectors which have standard normal distributions. The components of y and z are refered to the speaker and residual factors, respectively.

The channel-dependent supervector c, which represents channel effects in a speech, is supposed to be distributed according to

$$c = Ux, \quad (8)$$

TABLE I
DESCRIPTION OF THE DATA FOR ESTIMATING UBM AND JFA PARAMETERS

	MFCC	phase
UBM estimating for male	1 utterance of the each 100 speakers	
UBM estimating for female	1 utterance of the each 150 speakers	1 utterance of the each 100 speakers
JFA parameter training for male	960 utterances of the 100 speakers	
JFA parameter training for female	1526 utterances of the 150 speakers	

where U is a rectangular matrix of low rank, and x has standard normal distribution. The components of x are refered to the channel factors.

A detailed description of JFA can be refered by literature [10].

IV. Combination method and decision method

MFCCs use only the magnitude of the Fourier Transform in time-domain speech frames, that is, phase component is ignored. On the other hand, phase information ignores the magnitude of the Fourier Transform in time-domain speech frames. Therefore, in this paper, the JFA score based on MFCCs is combined with the JFA score based on phase information. When a combination of the two methods is used to identify the speaker, the score of the MFCC-based JFA is linearly coupled with that of the JFA based on phase information to produce a new score $Score_{comb}$ given by

$$Score_{comb} = (1-\alpha)Score_{MFCC} + \alpha Score_{phase}, \qquad (9)$$

where $Score_{MFCC}$ and $Score_{phase}$ are the score produced by MFCC-based speaker model and phase information-based speaker model, respectively, and α denotes the weighting coefficients, which are determined empirically. The combination score is then compared to the threshold in order to take the final decision.

V. Experiments

A. Experimental setup

The effect of phase information-based JFA for speaker verification under channel mismatch condition was evaluated on the NIST 2003 SRE database [11]. The NIST 2003 SRE database consists of recordings of 356 speakers (149 males and 207 females), recorded in multiple conditions which include six transmission methods (CDMA, LAND, GSM, TDMA, CELLULAR and UNK), multiple telephones, multiple places, etc. Almost all the data for every speaker were recorded by different environments. Therefore, this speaker verification task is very difficult. The NIST 2003 SRE database was divided into three parts, data for estimating UBM and JFA parameters, enrollment data and test data because our group don't have anything other database of the NIST SRE series but the NIST 2003 SRE database. Table I describes the data for estimating UBM and JFA parameters and Table II describes the details of the enrollment data and test data. Concretely speaking, 149 male speakers were divided into 100 speakers for estimating UBM and JFA parameters and 49 speakers for test, and 207 female speakers were divided into 150 speakers for estimating UBM and JFA parameters and 57 speakers for test. The test corpus consisted of 402 true trials and 402 × 48 false trials for males, and 523 true trials and 523×56 false trials for females, respectively. We used gender-dependent UBMs containing 1024 Gaussians for MFCC and 256 Gaussians for phase information, respectively.

TABLE II
DESCRIPTION OF ENROLLMENT AND TEST DATA

(a) enrollment data	
male	49 utterances of 49 speakers
female	57 utterances of 57 speakers
duration/utterance	about 2 minutes
(b) test data	
male	402 utterances of 49 speakers
female	523 utterances of 57 speakers
duration/utterance	15-45 seconds

TABLE III
THE NUMBER OF FACTORS FOR JFA MODEL (SPEAKER, CHANNEL, RESIDUAL)

	MFCC	phase
male	4, 20, 20	40, 6, 6
female	30, 20, 20	40, 40, 40

To verify robustness of phase information-based JFA for channel variability, the speaker verification system using JFA is compared with the system using traditional GMM-UBM in this paper. For GMM-UBM, GMMs containing 1024 Gaussians for MFCC and 256 Gaussians for phase information applying Maximum a posteriori (MAP) adaptation from gender-dependent UBMs were used. For JFA, the number of speaker factors, channel factors and residual factors are shown in Table III. Table IV shows conditions for the speech analysis.

We applied voice activity detection (VAD) for speech data. A frame is judged to be a speech frame if there is a segment put between long silence segments more than 200 ms. Under this condition, about 75% of all the frames were judged to be speech frames.

B. Experimental results

The equal error rates (EERs) for speaker verification using phase information-based GMM-UBM and JFA are given in Table V. Phase information-based JFA showed the improvement of EERs of 5.14% for male and 3.10% for female compared to the system based on GMM-UBM. The results show that phase information-based JFA has the moderate performance for speaker verification and the degradation of speaker verification performance using phase information caused by speaker and channel variability was mitigated partly. The EERs for speaker verification using MFCC-based GMM-UBM and JFA, using a combination of MFCC and phase information are given in Table VI. The combination of MFCC and phase information achieved a better result than MFCC-based JFA which was one of the standard methods for speaker verification. This indicated that the phase information has complementary nature with MFCC.

TABLE IV
CONDITIONS FOR SPEECH ANALYSIS

sampling frequency	8 kHz	
	MFCC	phase
window size	25 ms	16 ms
window shift	10 ms	5 ms
frequency range	all	60-700 Hz
dimensions	60 (19 MFCCs + power, their Δ and $\Delta\Delta$ coefficients	24 (12 sin and 12 cos components

TABLE V
EERS FOR SPEAKER VERIFICATION USING PHASE INFORMATION (%)

	GMM-UBM	JFA
male	24.57	19.43
female	17.86	14.76

These results show that phase information is effective for the speaker verification, even under channel mismatch condition. On the other hand, a previous study showed that phase information was not effective under channel mismatch and channel distortion conditions [9]. The reason is that the influence of channel mismatch for phase information is mitigated by using the channel variability robust JFA method. To verify this, we evaluated EERs for speaker verification under transmission mode matched and mismatch condition, respectively. We think that channel characteristics varies drastically between different transmission modes. Here, transmission mode matched condition means that the enrollment utterance and test utterance have same transmission mode, while transmission mode mismatch condition means that the enrollment utterance and test utterance have different transmission mode. The experimental result is given in Table VII. For both MFCC and phase information, the system based on JFA showed the improvement of EERs compared to the system based on GMM-UBM under transmission mode mismatch condition. From Table VII, it is obvious that the JFA can partly remove the influence of transmission mode mismatch for phase information, but the influence is still large. Based on these results, the more improvement of the results shown in Table V and VI are expected by normalizing phase information for each transmission mode.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we conducted the speaker verification using pseudo-pitch synchronous phase information under channel mismatch condition. To mitigate the influence of phase information under channel mismatch condition, a channel variability robust speaker verification method was applied. Phase information-based JFA showed the improvement of EERs of 5.14% for male and 3.10% for female compared to the traditional system based on GMM-UBM. We obtained the better result than only MFCC by combining MFCC and phase information.

Phase information shows the lower EERs under transmission mode matched condition while the higher EERs under transmission mode mismatch condition. As a result, in future work, we will try to normalize phase information for each transmission mode.

TABLE VI
EERS FOR SPEAKER VERIFICATION USING MFCC AND COMBINATION OF MFCC AND PHASE (%)

		GMM-UBM	JFA
MFCC	male	7.71	6.97
	female	7.38	3.44
MFCC+phase	male	7.68	6.72
	female	6.65	3.25

TABLE VII
EERS FOR SPEAKER VERIFICATION UNDER TRANSMISSION MODE MATCHED AND MISMATCH CONDITION (%)

		match		mismatch	
		GMM-UBM	JFA	GMM-UBM	JFA
MFCC	male	5.76	5.88	15.68	10.72
	female	6.21	3.17	13.46	5.73
phase	male	21.11	16.18	41.45	37.07
	female	11.22	10.91	42.06	34.48
MFCC+ phase	male	5.33	5.00	15.73	10.83
	female	5.25	3.20	12.50	5.72

REFERENCES

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

[2] 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.

[3] 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.

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

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

[6] 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.

[7] S. Nakagawa, L. Wang and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio, Speech and Language Processing*, Vol.20, No.4, pp.1085-1095, 2012.

[8] K. Shimada, K. Yamamoto and S. Nakagawa, "Speaker identification using pseudo pitch synchronized phase information in voiced sound," *Proc. on APSIPA ASC 2011*, CD-ROM, Xi'an, China, Oct 2011.

[9] L. Wang and S. Nakagawa, "Speaker identification/verification for reverberant speech using phase information," *Proc. of WESPAC 2009*, No.0130 (8pages), 2009.

[10] P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Trans. on Audio, Speech and Language*, Vol.16, No.5, pp.980-988, July 2008.

[11] "http://www.nist.gov/speech/tests/spk/2003/index .htm".