Relative phase information for detecting human speech and spoofed speech

Longbiao Wang[1], *Yohei Yoshida*[1], *Yuta Kawakami*[1] *and Seiichi Nakagawa*[2]

[1]Nagaoka University of Technology, Japan
[2]Toyohashi University of Technology, Japan
{wang@vos, s123182@stn, s123118@stn}.nagaokaut.ac.jp, nakagawa@slp.cs.tut.ac.jp

Abstract

The detection of human and spoofed (synthetic/converted) speech has started to receive more attention. In this study, relative phase information extracted from a Fourier spectrum is used to detect human and spoofed speech. Because original/natural phase information is almost entirely lost in spoofed speech using current synthesis/conversion techniques, a modified group delay based feature, the frequency derivative of the phase spectrum, has been shown effective for detecting human speech and spoofed speech. The modified group delay based phase contains both the magnitude spectrum and phase information. Therefore, the relative phase information, which contains only phase information, is expected to achieve a better spoofing detection performance. In this study, the relative phase information is also combined with the Mel-Frequency Cepstral Coefficient (MFCC) and modified group delay. The proposed method was evaluated using the "ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge" dataset. The results show that the proposed relative phase information significantly outperforms the MFCC and modified group delay. The equal error rate (EER) was reduced from 1.74% of MFCC, 0.83% of modified group delay to 0.013% of relative phase. By combining the relative phase with MFCC and modified group delay, the EER was reduced to 0.002%.

Index Terms: Spoofing detection, relative phase information, group delay, GMM, countermeasures

1. Introduction

Recently, speaker verification technology has been used in many applications using telephone, such as telephone banking and credit cards [1, 2]. However, the conventional speaker verification system is weak for voice conversion and speech synthesis techniques [3, 4]. In voice conversion, the speech of a source speaker is converted to voice like a target speaker. For speech synthesis, the voice of the target speaker is mimicked given any text. Related studies have indicated that the detection of spoofed speech (synthetic/converted speech) from human speech is very important to improve the robustness of speaker verification systems [5, 6, 7, 8, 9, 10]. In this study, we focus on spoofing detection, a task to determine whether a speech sample contains human or spoofed speech.

To detect spoofed speech from human speech, many features (e.g. magnitude spectrum, pitch, group delay and modulation features) have been considered [5, 9, 11]. In addition to pitch information, spectral information was proposed to detect synthetic speech [5]. In [11], cosine-normalized phase and modified group delay function phase spectrum based features were proposed to distinguish voice converted speech from human speech. In [9], modulation features were applied to detect synthetic speech. These studies indicate that phase related features outperform magnitude-based features because the original phase information is lost in the spoofed speech.

The most commonly used phase related feature may be the group delay based feature [13, 14]. Group delay is defined as the negative derivative of the phase of the Fourier transform of a signal. In fact, the group delay based phase contains both the magnitude spectrum and phase spectrum [12, 13, 14]. This means the component of magnitude spectrum in group delay may degrade the performance of spoofing detection. In our previous study [15, 16, 17, 18], relative phase information directly extracted from the Fourier transform of the speech wave has been proposed. To reduce the phase variation by cutting positions, the phase of a certain base frequency is kept constant, and the phases of other frequencies are estimated relative to this. The experimental results showed that the relative phase information was effective for speaker recognition for various conditions. In this paper, the relative phase information is proposed to detect human speech and spoofed speech. Because the relative phase information does not contain any magnitude spectrum and cannot normalize the phase variation by cutting positions, it is expected to achieve a better performance than other phase relative features such as the group delay based feature. Furthermore, the relative phase information is combined with modified group delay for spoofing detection.

The remainder of this paper is organized as follows: The system of spoofing detection is described in Section 2. Section 3 presents the modified group delay and the relative phase information extraction. The experimental setup and results are reported in Section 4, and Section 5 presents our conclusions.

2. Overview of spoofing detection system

The flowchart of the spoofing detection system is shown in Fig. 1. In this study, a Gaussian mixture model (GMM) [21, 22] is used as spoofed speech detector. The decision about whether speech is natural human or spoofed is based on the log likelihood ratio:

$$\Lambda(O) = \log p(O|\lambda_{human}) - \log p(O|\lambda_{spoof}), \quad (1)$$

where O is the feature vector of input speech, λ_{human} and λ_{spoof} are the GMMs for natural and spoofed speech, respectively. Here, Mel-frequency Cepstral Coefficient (MFCC), modified group delay and relative phase information described in Section 3 are used.

In this study, the likelihood ratios of two or three features are also linearly combined to produce a new score $\Lambda_{comb}(O)$ given by

$$\Lambda_{comb}(O) = \sum_n \alpha_n \Lambda(O_n), \quad (2)$$

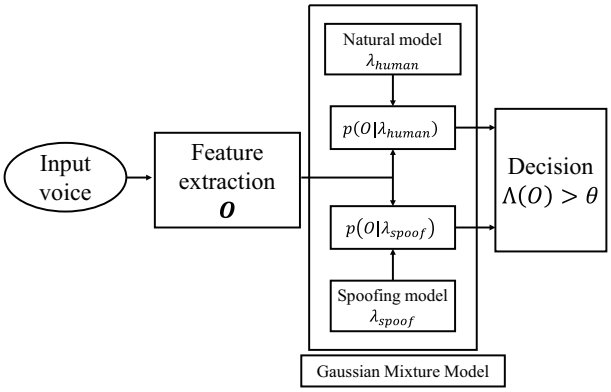

Figure 1: *Flowchart of spoofing detection system.*

where $\Lambda(O_n)$ is the log likelihood ratio and α_n denotes the weighting coefficients corresponding to the n-th feature set $n \in \{1, 2, 3\}$ is MFCCs, MGDCC or Relative phase, respectively. The decision threshold and weighting coefficient were determined using a development set.

3. Phase information extraction

3.1. Modified group delay

The spectrum $X(\omega)$ of a signal is obtained by DFT of an input speech signal sequence x(n)

$$X(\omega) = |X(\omega)|e^{j\theta(\omega)}, \tag{3}$$

where $|X(\omega)|$ and $\theta(\omega)$ are the magnitude spectrum and phase spectrum at frequency ω, respectively.

Group delay [23] is defined as the negative derivative of the Fourier transform phase for frequency, that is,

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega}. \tag{4}$$

The group delay function can also be calculated directly from the speech signal using

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2}, \tag{5}$$

where the subscripts R and I denote the real and imaginary parts of the Fourier transform. $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x(n)$ and $nx(n)$, respectively.

There are many studies reporting that modified group delay is better than the original group delay [12, 13, 14, 23]. The modified group delay function can be defined as

$$\tau_m(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S_c(\omega)}, \tag{6}$$

where $S_c(\omega)$ is the cepstrally smoothed spectrum of $S(\omega)$ and $S(\omega)$ is the squared magnitude $|X(\omega)|^2$ of the signal $x(n)$.

3.2. Relative phase information

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 base frequency ω is kept constant, and the phases of other frequencies are estimated relative to this. For example, by setting the base frequency ω to 0, we obtain

$$X'(\omega) = |X(\omega)| \times e^{j\theta(\omega)} \times e^{j(-\theta(\omega))}, \tag{7}$$

whereas for the other frequency $\omega' = 2\pi f'$, the spectrum becomes [18]

$$\begin{aligned} & X'(\omega') \\ = \quad & |X'(\omega')| \times e^{j\theta(\omega')} \times e^{j\frac{\omega'}{\omega}(-\theta(\omega))}. \end{aligned} \tag{8}$$

In this way, the phase can be normalized, and the normalized phase information becomes

$$\tilde{\theta}(\omega') \quad = \quad \theta(\omega') + \frac{\omega'}{\omega}(-\theta(\omega)). \tag{9}$$

In the experiments described in this paper, the base frequency ω is set to $2\pi\times$ 1000 Hz. In the previous study, we used phase information only in a sub-band frequency range to reduce the number of feature parameters. 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 is $\approx 2\pi$, despite the two phases being very similar to each other. Therefore, we modified the phase into coordinates on a unit circle [18], that is,

$$\tilde{\theta} \rightarrow \{\cos\tilde{\theta}, \sin\tilde{\theta}\}. \tag{10}$$

We can reduce the phase variation using the relative phase extraction method that normalizes the phase variation by cutting positions. 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 by $\frac{2\pi}{16}$, while for a 500-Hz periodic wave, the phase shifts only by $\frac{2\pi}{32}$ with this single sample cutting shift. However , if the 17 sample points shift, the phases of the 1000-Hz and 500-Hz waves will shift by $\frac{17\cdot 2\pi}{16} (\mathrm{mod}\ 2\pi) = \frac{2\pi}{16}$ and $\frac{34\pi}{32}$, respectively. Therefore, the values of the relative phase information for different cutting positions are very different from those of the original cutting position. The phase variation is summarized in Table 1. We have partly addressed such variations using a statistical GMM [18].

Table 1: *Phase variation related to the frequency ω and sample points Δ of shifted position.*

Period	Frequency	Phase variation
T	$\omega = \frac{2\pi}{T}$	$\frac{\Delta}{T}2\pi$

If we could split the utterance by each pitch cycle, changes in the phase information would be further obviated. Thus, we proposed a new extraction method that synchronizes the splitting section with a pseudo-pitch cycle [19, 20]. With respect to how to unite the cutting sections in the time domain, the proposed method looks for the maximum amplitude at the center of 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. 2 outlines how to synchronize the splitting section. We expect an improvement over our proposed conventional phase information [16, 17, 18].

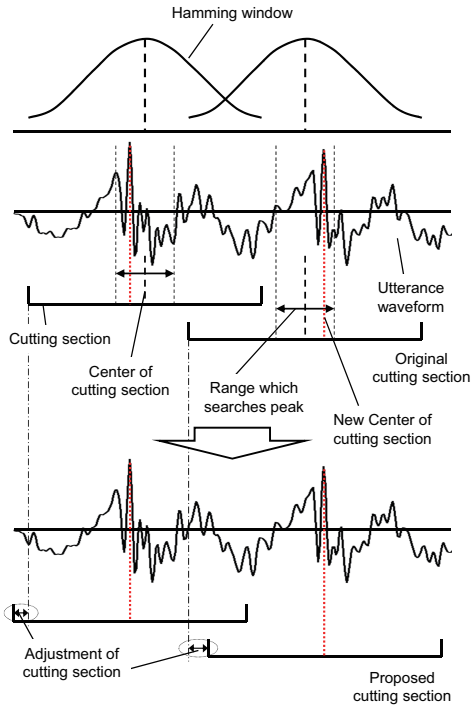

Figure 2: *How to synchronize the splitting section.*

4. Experiments

4.1. Datasets

We evaluate our proposed method for spoofing detection using the standard "ASVSpoof 2015 Challenge" dataset [1] of both genuine (human) and spoofed speech. Genuine speech was collected from 106 speakers (45 male, 61 female) with no significant channel or background noise effects. Spoofed speech was generated from the genuine data using a number of different spoofing algorithms. The full dataset was partitioned into three subsets, the first for training, the second for development and the third for evaluation. The details of each subset are summarized in Table 2. There was no speaker overlap across the three subsets regarding target speakers used in voice conversion or Text To Speech (TTS) adaptation.

For the training dataset, each spoofed utterance was generated according to one of three voice conversion and two speech synthesis algorithms. For the development dataset, spoofed speech was generated according to one of the same five spoofing algorithms used to generate the training dataset. For the evaluation dataset, spoofed data was generated according to diverse spoofing algorithms. They included the same five algorithms used to generate the development dataset in addition to others, designated as "unknown" spoofing algorithms.

4.2. Experimental setup

The input speech was sampled at 16 kHz. For MFCCs, a total of 38 dimensions (12 MFCCs, 12 ΔMFCCs, 12 $\Delta\Delta$MFCCs, Δpower and $\Delta\Delta$power) were calculated every 10 ms with a window of 25 ms. Thirty-eight static modified group delay cepstral coefficients (MGDCC) were calculated from the modified group delay function phase spectrum [9]. Relative phase information was calculated every 5 ms with a window of 12.5 ms. A spectrum with 128 components consisting of magnitude and phase was calculated by DFT for every 256 samples. Then 39 static relative phase features (that is, 19 $\cos\tilde{\theta}$ and 19 $\sin\tilde{\theta}$) were extracted. For the pseudo-pitch-synchronized phase information extraction method, the range for searching the peak amplitude point is 2.5 ms (half of the frame shift). The details of analysis conditions for MFCC, MGDCC and relative phase information are summarized in Table 3.

GMMs of human and spoofed speech were trained using a training dataset, and the mixed number of GMMs was 256, as determined by the development dataset.

[1] http://www.spoofingchallenge.org/

Table 2: *Number of non-overlapping target speakers and utterances in the training, development and evaluation datasets.*

Subset	#Speakers		#Utterances	
	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	$\approx$ 200000

Table 3: *Analysis conditions for MFCC, MGDCC and relative phase information.*

	MFCC	MGDCC	Relative phase
Frame length	25 ms	25 ms	12.5 ms
Frame shift	10 ms	10 ms	5 ms
FFT size	512 samples (400 data plus 112 zeros)	512 samples (400 data plus 112 zeros)	256 samples (200 data plus 56 zeros)
Dimensions	38		

4.3. Experimental results

4.3.1. Results of development dataset

The Equal Error Rates (EERs) of spoofing detection performance for the development dataset are shown in Table 4. The modified group delay cepstral coefficient (MGDCC) outperforms MFCC. The results show the same trend as [11]. Because the MGDCC also contains magnitude spectrum information, the spoofing detection performance is not sufficient. Relative phase information significantly outperforms the MGDCC because it normalizes the phase variation by cutting positions. The combination of relative phase with MFCC or MGDCC is also significantly better than the combination of MGDCC with

Table 4: *EERs (%) of spoofing detection performance of various features on development dataset.*

Features	Equal error rate (%)
MFCC	1.74
MGDCC	0.83
Relative phase	0.013
MFCC+MGDCC	0.256
MFCC+relative phase	0.004
MGDCC+relative phase	0.004
MFCC+MGDCC+relative phase	0.002

Table 5: *EERs (%) of spoofing detection performance of various features on evaluation dataset.*

Features	Know attacks						Unknown attacks						All attacks
	s1	s2	s3	s4	s5	Ave.	s6	s7	s8	s9	s10	Ave.	Ave.
MGDCC	—	—	—	—	—	1.155	—	—	—	—	—	6.761	3.958
Relative phase	0.000	0.025	0.000	0.000	0.025	0.010	0.285	0.005	1.179	0.000	37.728	7.840	3.925
MGDCC+ relative phase	0.000	0.009	0.000	0.000	0.015	0.005	0.081	0.005	0.080	0.000	37.068	7.447	3.726

MFCC. By combining the log likelihood ratios of three features (two phase related features and one magnitude related feature), a best performance is achieved, that is, the EER is from 0.256% of the combination of MGDCC with MFCC to 0.002% of the proposed method.

4.3.2. Results of evaluation dataset

The Equal Error Rates (EERs) of spoofing detection performance on evaluation dataset are shown in Table 5. Because we cannot submit MFCC based log likelihood ratio to "ASVSpoof 2015 Challenge" in time and we do not have a key file for the evaluation set, only the phase related results are available in this paper. For "known attacks", the trend of the evaluation dataset is the same as that of the development dataset. Our result of the combination of MGDCC and relative phase for "known attacks" submitted to "ASVSpoof 2015 Challenge" achieved 2nd place ranking among 16 teams even when using a very simple GMM based detector without any score normalization. For "unknown attacks", both phase related features achieved good performance except for "s10" spoofed speech. The reason may be that the phase related feature is weak for an unknown "s10" voice conversion or speech synthesis technique considering phase information. However, we do not have access to the detailed analysis as the key file for the evaluation dataset was unavailable. In the development dataset, the combination of MFCC with two phase related features achieved the best performance. It is considered that the performance of "known attacks" and "unknown attacks" may be improved when we combine three features. Furthermore, state-of-the-art speaker verification, such as i-vector based feature representation and probabilistic linear discriminant analysis (PLDA) based modeling [24], is also expected to improve the spoofing detection performance.

5. Conclusions

In this paper, the relative phase information was proposed for spoofing detection, and was also combined with the MFCC and modified group delay cepstral coefficient. The proposed method was evaluated with the "ASVspoof 2015 Challenge" dataset. The results indicated that the proposed relative phase information significantly outperformed the MFCC and MGDCC. For the development dataset, the EER was reduced from 1.74% of MFCC, 0.83% of MGDCC to 0.013% of the relative phase. By combining the relative phase with MFCC and MGDCC, the EER was reduced to 0.002%.

For the evaluation dataset, the combination of MGDCC and relative phase for "known attacks" submitted to "ASVSpoof 2015 Challenge" achieved 2nd place among 16 teams, even although we only used a very simple GMM based detector without any score normalization. For "unknown attacks", both phase related features achieved good performance except for "s10" spoofed speech. The reason may be that the phase related feature is weak for an unknown voice conversion or speech synthesis technique considering phase information.

In our future work, we will try to combine relative phase information with MGDCC and MFCC for an evaluation dataset. Furthermore, we will try to implement the state-of-the-art i-vector based feature representation and PLDA based modeling for spoofing detection [24].

6. References

[1] Joseph P Campbell Jr, "Speaker recognition: A tutorial", Proc. of the IEEE, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] T Kinnunen, HZ Li, "An overview of text-independent speaker recognition: from features to supervectors", Speech Communication, vol.52, No. 1, pp. 12–40, 2010.

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," Speech and Audio Processing, IEEE Transactions on, vol. 6, no. 2, pp. 131–142, 1998.

[4] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm", Audio, Speech, and Language Processing, IEEE Trans-actions on, vol. 17, no. 1, pp. 66–83, 2009.

[5] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spec-trum and pitch", in Proc. of ICSLP, pp. 302–305, 2000.

[6] P.L. De Leon, I. Hernaez, I. Saratxage, M. Pucher and J. Yamagishi, "Detection of synthetic speech for the problem of imposture", Proc. of ICASSP, pp. 4844-4847, 2011.

[7] P.L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech", Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 8, pp. 2280-2290, 2012.

[8] Q. Jin, A.R. Toth, A.W. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?", in Proc. of ICASSP, pp. 4845–4848, 2008.

[9] Z. Wu, X. Xiao, E. Chng, H. Li, "Synthetic speech detection using temporal modulation feature", Proc. of ICASSP, pp. 7234–7238, 2013.

[10] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li, "Spoofing and countermeasures for speaker verification: a survey", Speech Communication, Vol.66, pp. 130–153, 2015.

[11] Z. Wu, E.S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition", in Proc. of Interspeech, 2012.

[12] 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.

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

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

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

[16] 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.

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

[18] 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.

[19] Y. Kawakami, L. Wang and S. Nakagawa, “Speaker identification using pseudo pitch synchronized phase information in noisy environments”, Proc. APSIPA, 5 pages, 2013.

[20] Y. Kawakami, L. Wang A. Kai and S. Nakagawa, “Speaker Identification by Combining Various Vocal Tract and Vocal Source Features”, Proc. of International Conference on Text, Speech and Dialogue 2014, pp. 382-389, Sep. 2014.

[21] D A Reynolds, “Speaker identification and verification using Gaussian mixture speaker models”, Speech Communication, Vol. 17, No. 1–2, pp. 91–108, 1995.

[22] L. Wang, N. Kitaoka and S. Nakagawa, “Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM”, Speech Communication, Vol. 49, No.6, pp. 501-513, 2007.

[23] R. Hegde, H. Murthy and V. Gadde, “Significance of the modified group delay feature in speech recognition”, Audio, Speech and Language Processing, IEEE Transactions on, vol. 15, no. 1, pp. 190–202, 2007.

[24] Y. Jiang, K. Lee and L. Wang, “PLDA in the I-Supervector Space for Text-Independent Speaker Verification”, EURASIP Journal on Audio, Music and Speech Processing, 2014:29, 2014.