REPLAY ATTACK DETECTION USING MAGNITUDE AND PHASE INFORMATION WITH ATTENTION-BASED ADAPTIVE FILTERS

Meng Liu[1], *Longbiao Wang*[1,*], *Jianwu Dang*[1,2,*], *Seiichi Nakagawa*[3], *Haotian Guan*[4], *Xiangang Li*[5]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]Chubu University, Kasugai, Japan
[4]Huiyan Technology (Tianjin) Co. Ltd., Tianjin, China
[5]AI Labs, Didi Chuxing, Beijing, China

ABSTRACT

Automatic Speech Verification (ASV) systems are highly vulnerable to spoofing attacks, and replay attack poses the greatest threat among various spoofing attacks. In this paper, we propose a novel multi-channel feature extraction method with attention-based adaptive filters (AAF). Original phase information, discarded by conventional feature extraction techniques after Fast Fourier Transform (FFT), is promising in distinguishing genuine from replay spoofed speech. Accordingly, phase and magnitude information are respectively extracted as phase channel and magnitude channel complementary features in our system. First, we make discriminative ability analysis on full frequency bands with F-ratio methods. Then attention-based adaptive filters are implemented to maximize capturing of high discriminative information on frequency bands, and the results on ASVspoof 2017 challenge indicate that our proposed approach achieved relative error reduction rates of 78.7% and 59.8% on development and evaluation dataset than the baseline method.

Index Terms— replay attacks, phase information, frequency bands, adaptive filters, ASVspoof 2017

1. INTRODUCTION

Automatic Speech Verification (ASV) systems [1] are required to be robust against spoofing attacks, and detection of spoofed speeches (synthetic/converted/replay) has started to receive more attention [2]. Among various spoofing attacks, replay attack poses the greatest threat [3].

ASVspoof 2017 Challenge was organized with a focus on the limitations of existing preventive measures against replay attacks, offering participants Constant Q Cepstral Coefficients (CQCC) as a baseline feature set with a simple Gaussian Mixture Model (GMM) as a classifier. Most of the anti-spoofing systems currently proposed for replay attack detection applied magnitude-based features derived from speech signals. These include Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC) [4], Rectangular Frequency Cepstral Coefficients (RFCC) [5] and Frequency Domain Linear Prediction (FDLP) [6]. Although the magnitude-based features have proven to be effective and widely used, it is also important to explore the effect of the phase domain features, since there may be complementary channel information to the magnitude-based features and little related work has been done in replay attack detection.

Evidence of the effective utilization of the phase-related feature in spoof speech detection (SSD) systems can be found in related studies [7, 8]. For speech synthesis (SS) and voice conversion (VC) attacks, Relative Phase (RP) has been proposed with good performance [9]. Francis et al. utilized a heavy end-to-end deep learning framework with Group Delay gram (GD-gram) to detect replay attacks. As the best-performing system it implies promising effectiveness of the feature containing phase for this task [10].

Frequency modulation also contributes to replay detection. These include Inverted Mel-Frequency Cepstral Coefficients (IMFCC) [11], high-frequency feature [12], high-level feature from log power spectrum with the light convolution neural network (LCNN) as a back-end [13], etc.

The motivations behind this work are: (1) For replay attack detection, phase-based feature contains high discriminative information discarded by the magnitude-based feature, hence incorporation of these two channels' information would be beneficial; and (2) Since recording and playback scenes are complicated, features should be robust and adaptive rather than adhering to one invariable extracting method.

Thus in this paper, we propose attention-based adaptive filters to extract high discriminating phase features based on Relative Phase (RP) method [14]. We term the novel attention-based adaptive Relative Phase (ARP) for replay detection. This work also explores the use of the new proposed method for extracting magnitude-based features that we term

*Corresponding author: longbiao_wang@tju.edu.cn; jdang@jaist.ac.jp.

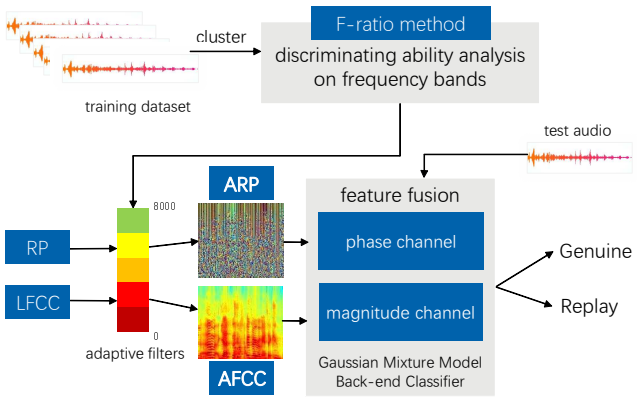


Fig. 1. Overview of the proposed framework for audio replay attack detection. (Note: RP: Relative Phase, LFCC: Linear Frequency Cepstral Coefficients).

attention-based Adaptive Frequency Cepstral Coefficients (AFCC). Furthermore, the new proposed features are also utilized as input of two channels for our replay detection system due to their complementarity. Our proposed method is compared with the CQCC (baseline method), MFCC, IMFCC, RP on ASVspoof 2017 Challenge database [15].

2. MAGNITUDE AND PHASE INFORMATION WITH ATTENTION-BASED ADAPTIVE FILTERS

As is illustrated in Fig. 1, our proposed framework could be described by three parts: (1) full frequency band analysis on discriminating abilities using F-ratio methods and the design of attention-based adaptive filters (AAF), (2) proposal of the novel adaptive relative phase (ARP) feature and adaptive frequency cepstral coefficient (AFCC) feature, (3) score fusion using phase and magnitude information to detect replay attacks. In the following subsections, these three major components will be described in detail, especially the proposed ARP feature.

2.1. Frequency Bands Analysis Using F-ratio

The F-ratio has been presented to improve performance of speaker recognition in early years, and it has shown effectiveness on emphasizing individual information (inter-speaker) and restraining linguistic information (intra-speaker) [16]. For anti-spoofing task, our goal is to enhance high discriminative information between genuine class and replay class and suppress speaker individual or linguistic information. Guided by this idea, we imported the classic F-ratio analysis method and implemented it on magnitude spectrum and phase spectrum. The $\boldsymbol{Fratio}$ is defined as:

$$\boldsymbol{Fratio} = \frac{\sum_{i=1}^{M}(\boldsymbol{u}_i - \boldsymbol{u})^2}{\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N}(\boldsymbol{x}_i^j - \boldsymbol{u}_i)^2}, \tag{1}$$

where $\boldsymbol{x}_i^j$ is the j_{th} sample feature vector of class i with $j = 1, 2, ..., N$, and $i = 1, 2$. $\boldsymbol{u}_i$ and $\boldsymbol{u}$ are the average vectors for class i and for all classes (genuine and replay) respectively, which are defined as:

$$\boldsymbol{u}_i = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{x}_i^j, \boldsymbol{u} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{u}_i. \tag{2}$$

The dimension of $\boldsymbol{Fratio}$ (d) corresponds with the sample vector in the genuine and replay classes:

$$d = dim(\boldsymbol{Fratio}) = dim(\boldsymbol{x}) = dim(\boldsymbol{u}) \tag{3}$$

Eq. (1) obtains a d-dimensional ratio vector between inter-class variance and intra-class variance. Then we can map the $\boldsymbol{Fratio}$ value to frequency axis uniformly, so that the discriminative ability on full frequency bands could be described by the $\boldsymbol{Fratio}$ value. $\boldsymbol{Fratio}_s$ denotes the $\boldsymbol{Fratio}$ value of the s_{th} frequency band point.

2.2. Attention-based Adaptive Filters (AAF)

For speech research, filters are designed to discard valueless information and keep informative parts according to the specific task [17]. MFCC utilizes a mel scale to imitate human hearing characteristics with a focus on low frequency regions. But for replay detection, substantial informative and discriminative information do not only exist in the auditory frequency range. Therefore, the adaptive scale is proposed in this study to pay more attention to the high discriminative frequency region adaptively rather than certain regions such as in IMFCC.

The proposed attention-based adaptive filters (AAF) could be described by the filter distribution density d_f. The whole frequency bands could be divided into several intervals by an attention threshold $\emptyset$. The frequency ranges above the attention threshold are regions containing high discriminative parts with a dense filter distribution, correspondingly, sparse in the ranges below a threshold. The threshold $\emptyset$ is set as:

$$\emptyset = \frac{1}{F * S}\sum_{f=1}^{F}\sum_{s=1}^{S}\boldsymbol{Fratio}_s^f, \tag{4}$$

where $\boldsymbol{Fratio}_s^f$ means $\boldsymbol{Fratio}$ value of the s_{th} band point in the f_{th} frequency range with $s = 1, 2, ..., S$ and $f = 1, 2, ..., F$. Filter distribution density d_f is defined as follows:

$$d_f = \frac{\sum_{s=1}^{S}\boldsymbol{Fratio}_s^f}{\sum_{f=1}^{F}\sum_{s=1}^{S}\boldsymbol{Fratio}_s^f}. \tag{5}$$

Then we could calculate filter numbers in each frequency range, and adaptive scale filterbanks could be constructed.

2.3. Proposed Adaptive Relative Phase (ARP) Using AAF

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

$$X(w) = |X(w)|e^{j\theta(w)}, \tag{6}$$

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

However, the phase changes depending on the clipping position of the input speech waveform even at the same frequency w. To overcome this problem, the phase of a certain base frequency w is kept constant, and phases of other frequencies are estimated relative to this. For example, by setting the base frequency w to θ, we obtain:

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

whereas for the other frequency $w' = 2f'$, the spectrum becomes:

$$X'(w)' = |X'(w')| \times e^{j\theta}(w') \times e^{j\frac{w'}{w}(-\theta(w))}. \tag{8}$$

After normalization, the phase $\theta(w')$ is transformed to:

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

Then the phase is mapped into the coordinates on a unit circle so that $\tilde{\theta}$ could be constraint to $\{cos\tilde{\theta}, sin\tilde{\theta}\}$. In this study, the relative phase is converted to the adaptive scale using attention-based adaptive filter to improve the classification quality, shown as follows:

$$RP(w') \xrightarrow{*AdaptiveFilterbank} ARP(w'') \tag{10}$$

2.4. Feature Complementarity Using Magnitude and Phase Information

In this study, we utilize a GMM-based replay speech detector with input of phase-channel and magnitude-channel information, and take Eq. (11) as measurements:

$$Score(O) = \log p(O|\lambda_g) - \log p(O|\lambda_s), \tag{11}$$

where O is the feature vector of input speech, λ_g and λ_s are the GMMs for original and replay speech, respectively.

For the score level fusion, we applied the linear combination proposed in [18] to obtain the final decision L:

$$L = \alpha L_1 + (1 - \alpha)L_2, \tag{12}$$

$$\alpha = \frac{\bar{L_1}}{\bar{L_1} + \bar{L_2}}, \tag{13}$$

where L_1 and L_2 represent scores from independent feature extracting models, L_1 and L_2 denote the averaged L_1 and L_2 over all the training data, respectively.

Table 1. Details of ASVspoof 2017 datasets.

Dataset	Number of speakers	Utterances	
		genuine	spoof
Training	10	1508	1508
Development	8	760	950
Evaluation	24	1298	12008

3. EXPERIMENTS

3.1. Database

The ASVspoof 2017 challenge database originates from the RedDots corpus [19], which is collected by ASV researchers worldwide under various environments and unseen scenarios. Three datasets are involved: train, development and evaluation. The sampling rate is set at 16 kHz with sample precision of 16 bits. Details of the database are shown in Table 1.

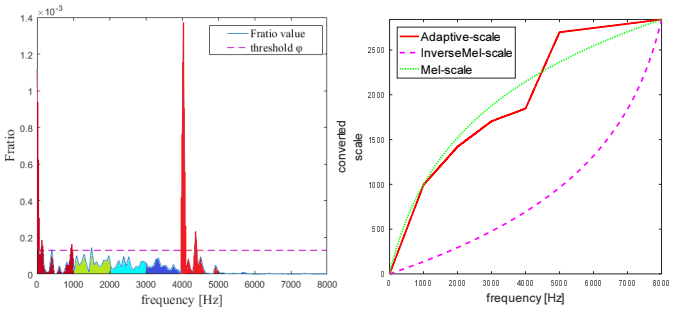
3.2. Experimental Setup

For F-ratio analysis on phase channel, the relative phase is calculated every 5 ms with a window of 12.5 ms. A series of experiments are conducted, and 118-dimensional static relative phase features (that is, 59 $cos\theta$ and 59 $cos\theta$) have the best tuning performance. For the magnitude channel, 39-dimensional LFCC (13 LFCC, 13Δ LFCC, 13$\Delta\Delta$LFCC) is chosen.

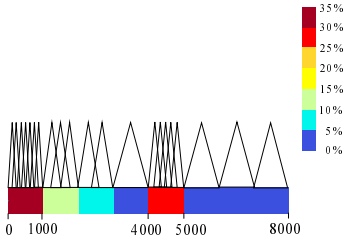
For the front-end feature extraction, CQCC is obtained by a default setting of 96 bins-per-octave and 16 uniform samples in the first octave. All relative phase related features adopt a number of 118 as dimensions. MelRP feature is calculated through a mel-scale filter, and ARP is filtered by our self-designed AAF. The magnitude-based feature in this study keeps the same setting as the phase feature. Two back-end classifiers, Gaussian Mixture Models of 512 components, are trained using the EM (Expectation Maximization) algorithm, on genuine and spoof utterances.

3.3. Results and Discussion

Based on the relative phase extraction method, a frequency band analysis of discriminative ability has been made on training and development dataset. Fig. 2 indicates that the ranges [0, 1000 Hz] and [4000 Hz, 5000 Hz] are more informative and discriminative when the filters are densely placed. In comparison, the number of filters in other frequency regions should be determined by the sum of F-ratio values in that region. Fig. 2 (c) illustrates our self-designed adaptive filters and its frequency distribution characteristics. First, experiments are conducted with individual features based on the GMM detector. Results are summarized in Table 2. From Table 2, we can find that our proposed ARP and AFCC



(a) F-ratio Analysis. (b) Proposed adaptive-scale.



(c) Filter distribution density on frequency bands.

Fig. 2. Process of designing AAF based on relative phase method.

features outperform the CQCC, MFCC and MelRP features. Compared with CQCC, AFCC achieved 61.2% relative error reduction rate on development dataset, and ARP obtained 55.6% relative error reduction rate on evaluation dataset, which shows that the adaptive scale is better than multi-scale CQCC applied.The IMFCC resulted in a poor performance confronted with a dataset mismatch problem. The traditional phase method Modified Group Delay Cesptral Coefficients (MGDCC) [20] feature seems unsuitable for this task because of the equal attention to all frequency bands. Compared with MelRP, the proposed ARP feature has improved 3.38% for an absolute error reduction rate, which indicates that our attention mechanism is feasible and effective. The table also suggests that the phase-domain features investigated are more

Table 2. EERs % of spoofing detection performance of individual features.

Feature	Development	Evaluation
CQCC	10.35	28.48
MFCC	13.74	34.39
IMFCC	4.83	28.59
MGDCC	25.92	38.10
RP	19.86	25.68
MelRP	10.36	16.03
AFCC	**4.01**	27.80
ARP	9.11	**12.65**

Table 3. EERs % of spoofing detection performance of score fusion.

Feature	Development	Evaluation
CQCC+MFCC	10.75	29.33
CQCC+AFCC	3.57	28.02
CQCC+RP	9.06	20.98
CQCC+MelRP	5.02	13.88
CQCC+ARP	2.26	12.58
AFCC+ARP	2.23	11.95
ARP+AFCC+CQCC	2.20	11.43

robust on the evaluation dataset recorded under varied unseen situations. Then phase feature and magnitude feature were incorporated, and the results are shown in Table 3. Compared with the baseline system, our proposed AFCC+ARP achieved relative error reduction rates of 78.4% and 58.0% on the development and evaluation dataset, respectively. This may be attributed to the fact that the phase information is incorporated into the score decision system and is complementary to the magnitude features. Also, the adaptive filters with more attentions on high discriminative frequency regions contribute to this result. Our best result is obtained by ARP+AFCC+CQCC with an EER of 2.20% on the development dataset and 11.43% on the evaluation dataset.

4. CONCLUSION

In this study, we proposed the attention-based adaptive filters (AAF) after the full frequency band discriminative contribution analysis using the F-ratio method.A novel adaptive relative phase feature was proposed with the self-designed AAF, and accordingly, we obtained the AFCC feature in the magnitude domain. Finally, a replay detecting system using phase and magnitude complementary information was implemented. From practical feature extraction standpoint, this system is less complex compared to the state-of-the-art systems, with improvements of 78.7% and 59.8% on the development and evaluation datasets, respectively. The results showed effectiveness that the frequency regions with a higher discriminating capability should be more emphasized adaptively. Also, it is indicated that the phase-based feature is complementary to the magnitude-based feature, and our proposed ARP method seems more robust to complicated recording situations.

5. ACKNOWLEGEMENTS

The research was supported partially by the Tianjin Municipal Science and Technology Project (Grant No.18ZXZNGX00330), National Natural Science Foundation of China (No.61771333) and JSPS KAKENHI Grant (No.16K12461).

6. REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanili, Mohammed Sahidullah, Aleksandr Sizov, Nicholas Evans, and Massimiliano Todisco, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[3] Tomi Kinnunen, Md. Sahidullah, Hctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech 2017*, 2017, pp. 2–6.

[4] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," in *Proc. Interspeech 2017, Stockholm, Sweden*, 2017, pp. 27–31.

[5] C. Hanili, "Features and classifiers for replay spoofing attack detection," in *2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*, Nov 2017, pp. 1187–1191.

[6] B. Wickramasinghe, S. Irtza, E. Ambikairajah, and J. Epps, "Frequency domain linear prediction features for replay spoofing attack detection," in *Proc. Interspeech 2018*, 2018, pp. 661–665.

[7] Phillip L. De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *IEEE Transactions on Audio Speech & Language Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

[8] Jon Sanchez, Ibon Saratxaga, Inma Hernez, Eva Navas, Daniel Erro, and Tuomo Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Transactions on Information Forensics & Security*, vol. 10, no. 4, pp. 810–820, 2015.

[9] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 2092–2096.

[10] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," in *Proc. Interspeech 2018*, 2018, pp. 681–685.

[11] L. Li, Y. Chen, D. Wang, and T.F. Zheng, "A study on replay attack and anti-spoofing for automatic speaker verification," in *Proc. Interspeech 2017*, 2017, pp. 92–96.

[12] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," in *Proc. Interspeech 2017, Stockholm, Sweden*, 2017, pp. 27–31.

[13] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Proc. Interspeech 2017*, 2017, pp. 82–86.

[14] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, "Spoofing speech detection using modified relative phase information," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 4, pp. 660–670, 2017.

[15] D. Li, L. Wang, J. Dang, M. Liu, Z. Oo, S. Nakagawa, H. Guan, and X. Li, "Multiple phase information combination for replay attacks detection," in *Proc. Interspeech 2018*, 2018, pp. 656–660.

[16] X. Lu and J. Dang, "Physiological feature extraction for text independent speaker identification using non-uniform subband processing," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, vol. IV, pp. 461–464.

[17] M. Kamble, H. Tak, and H. Patil, "Effectiveness of speech demodulation-based features for replay detection," in *Proc. Interspeech 2018*, 2018, pp. 641–645.

[18] K. Phapatanaburi, L. Wang, Z. Oo, W. Li, S. Nakagawa, and M. Iwahashi, "Noise robust voice activity detection using joint phase and magnitude based feature enhancement," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 6, pp. 845–859, 2017.

[19] Tomi Kinnunen, Md Sahidullah, Mauro Falcone, Luca Costantini, and Kong Aik Lee, "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5395–5399.

[20] Rajesh M. Hegde, Hema A. Murthy, and V. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio Speech & Language Processing*, vol. 15, no. 1, pp. 190–202, 2006.