Exploiting Magnitude and Phase Aware Deep Neural Network for Replay Attack Detection

Khomdet Phapatanaburi[*†], **Prawit Buayai**[**], **Watcharaphon Naktong**[*], and **Jakkree Srinonchat**[***], Non-members

ABSTRACT

Magnitude and phase aware deep neural network (MP-aware DNN) based on Fast Fourier Transform information, has recently received more attention in many speech applications. However, little attention has been paid to its use in terms of replay attack detection developed for automatic speaker verification and countermeasures (ASVspoof 2017). This paper aims to investigate MP-aware DNN as a speech classification to detect a non-replayed (genuine) and a replayed speech. In order to exploit the advantage of the complementary classifier-based methods to improve the reliable detection decision, a novel method was proposed; combining MP-aware DNN with a standard replay attack detection (that is, the use of constant Q transform cepstral coefficients-based Gaussian mixture model classification: CQCC-based GMM). Experiments were evaluated using ASVspoof 2017 and a standard measure of detection performance called equal error rate (EER). The results showed that MP-aware DNN-based detection outperformed conventional DNN using only the magnitude feature. Moreover, we found that the score combination of CQCC-based GMM with MP-aware DNN achieved an additional improvement, indicating that MP-aware DNN could be very useful, especially when combined with CQCC-based GMM for replay attack detection.

Keywords: Phase Information, Magnitude and Phase Aware Deep Neural Network, Replay Attack Detection, ASVspoof 2017

1. INTRODUCTION

Automatic speaker verification (ASV) [1–4] is a recent active research topic in the speech community because it has reached the point of mass-market adoption such as credit cards, telephone banking, access control in smartphones, etc. However, conventional ASV system [5] is well known to be vulnerable to audio replay attacks. This problem limits the deployment of ASV application because the replay attacks [6–8] in pre-recorded speech samples of a target genuine speaker is used by an attacker to deceive an ASV system. Related studies have indicated that the detection of replayed speech and genuine speech is crucial to improve the robustness of ASV. In this paper, we concentrate on a task that determines whether a speech sample contains genuine or replayed speech, called replay attack detection.

The typical replay attack detection systems [6, 9–10] are composed of two modules: feature extraction (front-end) and classifier/detection (back-end). In the feature extraction module, the role of the process is to extract a sequence of short-term spectral vectors from a given speech waveform. In the classifier/detection, the sequences of short-term spectral vectors are matched with the trained models, and finally, the matched scores are compared with genuine/replayed speech threshold. Almost all of the conventional replay attack detection systems are based on magnitude-based feature extractions derived from fast Fourier Transform (FFT). In [9–10], the authors used Mel-frequency cepstral coefficient (MFCC) as a magnitude based feature. The results confirmed that Gaussian mixture model-based acoustic modeling approach using the MFCC serves as an effective detection against several kinds of replay attack that would defeat a conventional ASV verification system. Similarly, the use of pitch and MFCC based on a support vector machine (SVM) was studied in [6] to investigate the performance of false acceptances when the use of replayed speech opened by an attacker tries to deceive an ASV system. The authors of [10] used spectral bitmaps to detect replay attacks presented for a text-dependent ASV system. All these studies have used only magnitude information based on FFT and have discarded phase information which is half of the original speech signal. In this paper, to avoid phase information loss, the use of joint phase and magnitude-based features will be focused on, to dis-

Manuscript received on March 23, 2019 ; revised on December 13, 2019 ; accepted on December 18, 2019. This paper was recommended by Associate Editor Pornchai Phukpattaranont.

*The authors are with the Department of Telecommunication Engineering, Faculty of Engineering and Architecture, Rajamangala University of Technology Isan, Nakhonrachasrima, Thailand.

**The author is with the Department of Computer Science and Engineering, University of Yamanashi, Kofu, Japan.

***The author is with the Department of Electronics and Telecommunication Engineering, Faculty of Engineering Rajamangala University of Technology Thanyaburi, Thailand.

†Corresponding author. E-mail: khomdet.ph@rmuti.ac.th

Digital Object Identifier 10.37936/ecti-eec.2020182.240341

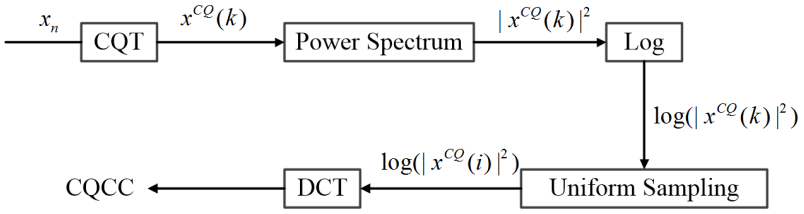

Fig.1: *Diagram of CQCC extraction process.*

tinguish between voice replayed speech and genuine speech.

Recently, to exploit full information in an original speech signal, magnitude and phase-aware deep neural network (MP-aware DNN), has been studied in many speech applications. This method uses the advantage of DNN which has no limitation on high feature dimensions and different features, which cannot normally be modeled by conventional shallow learning machine-based classification [12] such as GMM and SVM-based detection. Several studies have invested great effort in exploiting MP-aware DNN for speech regression and speech classification tasks. For instance, in [13], MP-aware DNN was proposed as a speech regression for enhancing conventional magnitude and phase feature in noise robust speaker recognition. The result showed that the proposed method augmenting magnitude (MFCC) and phase (modified group delay cepstral coefficient: MGDCC) feature could improve the speaker recognition rates when compared to DNN-based feature enhancement using only magnitude based feature. Similarly, the authors of [14–15] also applied MP-aware DNN to enhance the magnitude and phase of noisy speech by operating in the complex domain. In [16], MP-aware DNN was proposed as a speech classification based on noise robust voice activity detection. The result indicated that the MP-aware DNN could reduce the equal error rate (EER) of the VAD due to additional phase introduction in the DNN training. Although MP-aware DNN approaches have been applied in many speech applications, little attention has been paid to replay attack detection. Thus, this paper focuses on investigating the MP-aware DNN performance for replay attack detection.

In this paper, an MP-aware DNN is applied as a classification for the sake of distinguishing between genuine and replayed speech. MFCC and MGDCC were used as magnitude and phase information for the DNN input. The applied MP-aware DNN-based detection was compared with conventional DNN-based detection and baseline constant Q transform cepstral coefficients-based Gaussian mixture classification (CQCC-based GMM). Although the MP-aware DNN-based method may provide better discrimination detection than conventional DNN-based detection using only magnitude/phase information, it may not work well due to the lack of feature-based resolution within low frequencies [17] and limited training data [18]. To address this problem, a novel method was proposed; combining MP-aware DNN with CQCC-based GMM to improve the reliable detection decision. These two detections use different classifiers, features and may have a complementary effect based on different detection decision, we therefore expect the proposed combination of these methods to achieve a better performance compared with each individual method.

The remainder of this paper is organized as follows: Section 2 describes baseline replay attack detection used in recent works. Section 3 introduces magnitude and phase aware DNN (MP-aware DNN)-based detection. Section 4 describes the proposed combination of GMM and MP-aware DNN. The experimental setup and results for replay attack detection are evaluated in Section 5. Finally, Section 6 summarizes the paper and describes future work.

2. BASELINE REPLAY ATTACK DETECTION

This section introduces baseline replay attack detection based on CQCC and GMM, which are feature and detection, respectively. The details are described as follows.

2.1 CQCC

In this paper, CQCC is used as an input feature in baseline GMM-based detection as suggested in standard ASVspoof 2017 challenge. CQCC is a magnitude-based feature which jointly uses Constant Q Transform (CQT) with traditional cepstral analysis to capture speech signal. This mentioned feature is aimed at transforming geometric space of frequency bins to a linear space performing a frequency scale of the CQT, followed by a uniform resampling and a Discrete Cosine Transform (DCT). The extraction framework process is illustrated in Fig. 1. More details of CQCC feature extraction can be found in [17].

2.2 GMM-based detection

In this work, GMM was applied as the baseline replay attack detection : the available program was provided by the organizers of ASVspoof 2017 challenge. The flowchart of the GMM-based detection system is illustrated in Fig. 2. 512 components were

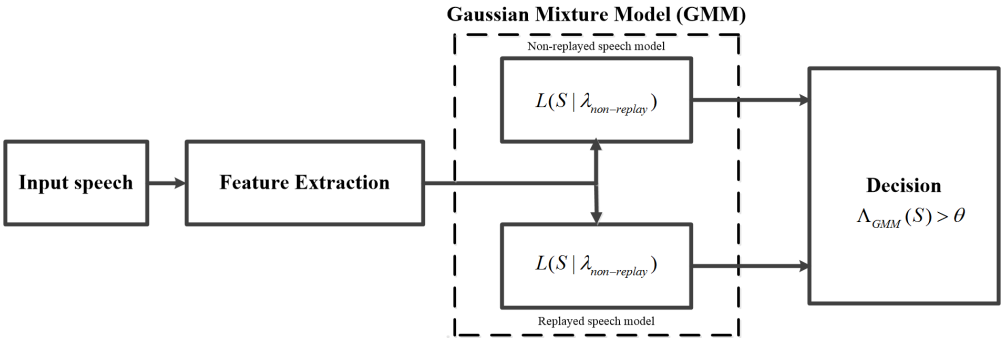

Fig.2: *Flowchart of GMM-based detection system.*

used. Two distinct models were learned for replayed and genuine speech with expectation maximisation (EM) algorithm with random initialisation. The decision about whether speech is replayed or genuine is based on the log likelihood ratio by using Eq. (1).

$$\wedge_{GMM}(S) = \log L(S|\lambda_{genuine}) - \log L(S|\lambda_{replay}) \tag{1}$$

where S is a sequence of feature vectors, L denotes the likelihood function, $\lambda_{genuine}$ and λ_{replay} represent GMM for genuine and replayed speech. In this paper, CQCC as described in the previous subsection was used as the feature input.

3. MP-AWARE DNN BASED DETECTION

This section introduces conventional and MP-aware DNN-based detection and two features used as DNN feature input. The details are described as follows.

3.1 Conventional and MP-aware DNN-based detection

DNN serves as a universal mapping function that could be used for both multi frame dependent classification and regression problem [19, 21]. Many studies [19–21] have shown that DNN is efficient for speech classification. In this paper, DNN was also applied as a discriminative detection for calculating the posterior probability of genuine and replay speech class from the given input feature explained in Section 3.2. A stochastic gradient descent (SGD), an optimization algorithm that improves cross-entropy error function, was used. The output layer comprises two neurons with softmax activation function that refers to the posterior probability of genuine and replayed speech class. The output score for a given feature is as follows:

$$\wedge_{DNN}(S) = P(\phi_{genuine}|S) - P(\phi_{replay}|S) \tag{2}$$

where $P(\phi_{genuine}|S)$ and $P(\phi_{replay}|S)$ are estimations of the posterior probability of genuine and replayed speech class, respectively. S is a sequence of feature vectors based on FFT information. For conventional DNN-based detection, only magnitude based feature is dependently used as the vectors of input speech as follows:

$$S = F_{Mag} \tag{3}$$

where F_{mag} uses only magnitude information. Here, MFCC briefly described in Section 3.2 is used. The structure of the DNN detection system is shown in Fig. 3(a). As observed in conventional DNN-based detection, phase information, half of the information present in the original signal, is ignored. Consequently, it may make the classification inefficient if only magnitude information-based feature is used.

Recently, phase information has been proven to be crucial for many speech processing tasks [22]. In [13, 16] DNN using phase features augmented with corresponding magnitude features was proposed as a multi-frame-dependent regression and classification task that could improve the performance of DNN using only magnitude feature. This can be seen in the simultaneous improvement of joint phase and magnitude feature. Here, there is an expectation that phase information can also enhance the performance of DNN training for a replay attack classification. Thus, the application of magnitude and phase information aware DNN, called MP-aware DNN, is used to determine whether a speech sample contains genuine or replayed speech. The structure of the MP-aware DNN is shown in Fig. 3(b). The feature vector of input speech, covering magnitude and phase information, is used as follows,

$$S = [F_{Mag},\ F_{Phase}] \tag{4}$$

where $Fphase$ is phase based features. By augmenting the phase with magnitude-based feature, DNN networks were trained and tested using augmenting MFCC and MGDCC feature.

3.2 Features used as DNN feature input

This subsection briefly introduces two features for conventional and MP-aware DNN-based detection,

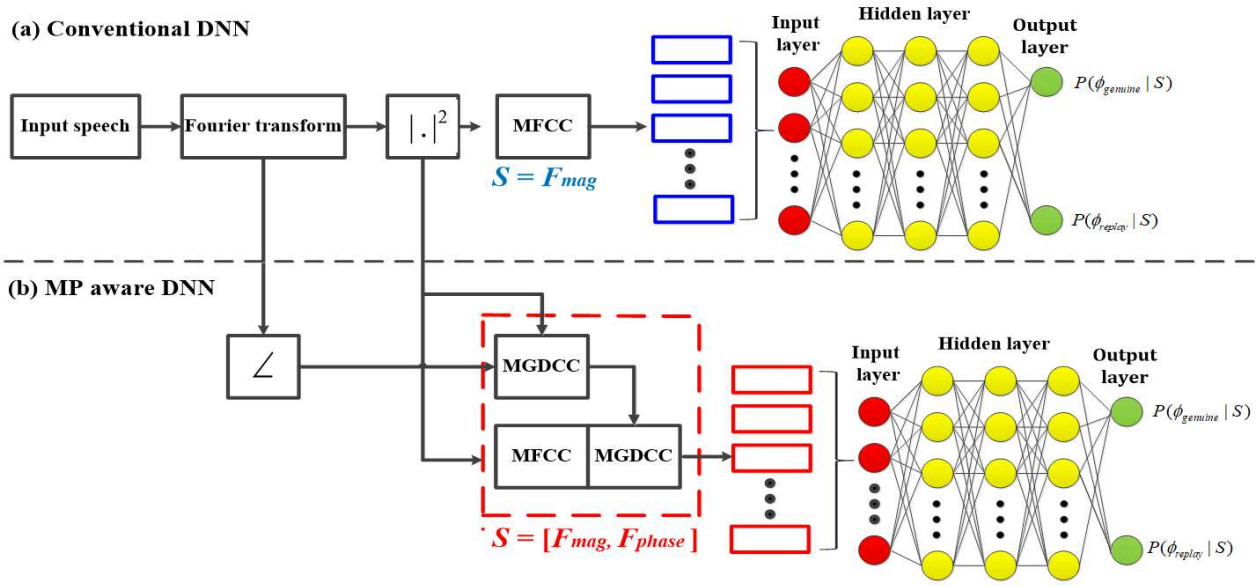

Fig.3: *Flowchart of two DNN-based detection systems where (a) is based on conventional DNN-based approach and (b) is MP-aware DNN-based approach.*

namely MFCC and MGDCC. The details of each technique are described as follows.

3.2.1 MFCC

MFCC is a popular magnitude-based feature for many speech applications including speaker verification and speech recognition. In this paper, MFCC is utilized as a magnitude-based feature for DNN input for the researcher to use the information and a reliable means of detecting replay attack [11].

3.2.2 MGDCC

Group delay spectrum has been exploited in automatic speech recognition for many years. It contains both magnitude and phase information, which is useful for replay attack detection. Group delay $\tau_x(\omega)$ is defined as the frequency differential of the phase spectrum:

$$\tau_x(\omega) = -\frac{d}{d\omega}\angle X(\omega) \tag{5}$$

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \tag{6}$$

Here, $X(\omega)$ is the Fourier transform of the signal $x(n)$, $Y(\omega)$ denotes the Fourier transform of $nx(n)$, footnote R and I indicate the real and imaginary parts of the complex. From the denominator of Eq. (6), we can receive an infinite value in the computation if the value of $|X(\omega)|$ is approximated to zero. Therefore, to avoid an explosion, modified group delay is smoothed as,

$$\tau_m(\omega) = \frac{\tau(\omega)}{|\tau(\omega)|}|\tau(\omega)|^{\alpha} \tag{7}$$

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}} \tag{8}$$

Here, $S(\omega)$ is cepstrally smoothed $X(\omega)$. The range of α and γ are $(0 < \alpha \leq 1.0)$ and $(0 < \gamma \leq 1.0)$. In this study, we used $\alpha = 0.4, \gamma = 0.9$ proposed by [23]. Moreover, to use some form of homomorphic processing, discrete cosine transform (DCT) is applied to convert $\tau(\omega)$ into meaningful and useful features for replay attack detection. The final feature is referred to as MGDCC.

4. PROPOSED COMBINATION OF GMM AND MP-AWARE DNN

In the previous section, MP-aware DNN method was introduced for replay attack detection. Although MP-aware DNN-based detection may provide a better performance than that of conventional DNN-based detection using only magnitude information, it may not work well due to the lack of feature-based resolution within low frequencies [17] and limited training data. In [18], a combination of GMM and DNN for distant-talking accent recognition was proposed to increase the accent recognition performance on limited training data. The result showed that the combination of these two different methods could improve the distant-accent recognition performance when compared to just an individual one. Motivated by [18], a combination of CQCC-based GMM and MP-aware DNN was proposed to take advantage of these benefits of different classifications and features and was

expected to achieve better performance. The probabilities obtained from the different systems are combined by the following equation.

$$\wedge_{comb}(S) = (1-\alpha)\wedge_{GMM}(S) - \alpha\wedge_{DNN}(S), \quad (9)$$
$$\alpha = \frac{\wedge_{GMM}(S)}{\wedge_{GMM}(S) + \wedge_{DNN}(S)}$$

where $\wedge_{GMM}$ and $\wedge_{DNN}$ are the score products of GMM and MP-aware DNN models, respectively. α is the weighing coefficient. The decision about whether a speech is genuine or replayed is computed using the maximum scores.

5. EXPERIMENTS

5.1 Experimental setup

To evaluate MP-aware DNN and the proposed method, experiments were conducted using ASVspoof 2017 database derived from RedDots corpus [24] where the original corpus served as the genuine recording and its replayed parts acted as spoofed recording. The database was categorized into three subsets including training (Train), development (Dev) and evaluation (Eval) subsets. The details of each subsection are shown in Table 1. For training all the detection methods introduced in Sections 2 and 3 were carried out using only the train subset, which contained 1508 non-replayed and corresponding replayed utterances derived from 10 speakers. For testing, the EER, which corresponded to the operating point with equal miss and false alarm rates, was implemented as a measure of detection performance. The available implementation provided by Bosaris Toolkit[1] was applied to compute the EER. The testing results were investigated using the Dev and the Eval subsets, respectively. The Dev subset was composed of 760 non-replayed and 950 replayed utterances with 8 speakers. The Eval subset contained 1298 non-replayed and 12008 replayed utterances with 24 speakers.

For baseline replay attack detection, GMM-based detection was done using the program of the organizers of ASVspoof 2017. Models of 512 components were used. Two distinct models were learned for replayed speech and non- replayed speech with expectation maximisation (EM) algorithm with random initialisation. 90-dimensional CQCC coefficients were used for input feature of GMM-based detection.

For conventional and MP-aware DNN-based detection, DNN were trained with SignalGraph Tool[2]. Here, the DNN setup of spoofing attack detection was followed as in [25] and the same parameters across different features were used. DNN took nine frames of the MFCC/MGDCC vectors in order to consider the correlation of adjacent frames. To avoid overfitting, weight decay and learning rate were set to 0.5 and 0.02, respectively. Network training was stopped when the error on the validation data started to increase. The detail of and model parameters for DNNs are shown in Table 2. Next, the analysis conditions of MFCC and MGDCC feature are shown in Table 3.

[1] https://sites.google.com/site/bosaristoolkit/home
[2] https://github.com/singaxiong/SignalGraph

Table 1: *The detail of ASVspoof 2017 database.*

Subsets	**#Speakers**	**# Utterances**	
		Genuine	Replayed
Train	10	1508	1508
Dev	8	760	950
Eval	24	1298	12008

Table 2: *DNN model parameters.*

Parameter name	**Value**
Batch size	256
Hidden node type	Sigmoid
Error function	Cross entropy
Training	Stochastic gradient decent
Learning rate	0.02
Weight decay	0.5
Context size	9

5.2 Experimental results

5.2.1 Result on baseline replay attack detection

This subsection shows the result of GMM-based detection using CQCC as the baseline performance. This result will be used to compare the proposed methods. The experimental results are shown in Table 4.

As observed in Table 4, the EER of Dev and Eval are 7.61 % and 28.09 %, respectively. This indicates that the Eval dataset is more difficult than the Dev dataset and that the baseline system still requires better improvement.

5.2.2 Results based on conventional and MP-aware DNN-based detection

In this subsection, the results based on conventional and MP-aware DNN-based detection was investigated. Before using the DNN-based model, different configurations of DNN based on MFCC feature were investigated to obtain the appropriate DNN for replay attack detection. DNNs were trained using the same parameters across different feature extraction as shown in Table 2. The number of layers varied from 2 to 3 and the number of nodes in each hidden layer increased from 256 to 1024. The results of DNN with two and three hidden layers with different hidden nodes are shown in Table 5.

From Table 5, based on MFCC, it can be seen that DNN using two layers with 256 hidden nodes

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

	MFCC	**MGDCC**
Frame length (ms)	20	20
Frame shift (ms)	10	10
FFT size (points)	512	512
Dimension	39 (13MFCCs, 13△s, and 13△△s)	36 (12MGDCCs, 12△s, and 12△△s)

Table 4: *Experimental results of GMM-based detection.*

Classifiers	**Features**	**EER (%)**	
		Dev	Eval
GMM	CQCC	7.61	28.09

achieved the best EER among all configurations. It seemed that the increasing number of the hidden layers could not provide consistent performance gain for replay attack detection. This might be because the architecture of DNN with only two-dimensional output could not handle the training data well as seen in [26].

Based on MFCC, the DNN using two layers with 256 hidden nodes was considered to be the best EER. To use the same standard DNN network configuration in our work, DNNs with remaining features were also trained using two layers with 256 hidden nodes.

In the same configurations as seen in Table 5, it can be seen that the MP-aware DNN using joint magnitude and phase information (MFCC+MGDCC) gave better performance than the DNN using the single feature (MFCC/MGDCC). This is due to the fact that joint magnitude and phase information make the DNN training/testing more efficient. The tendency is similar to a speaker recognition task referring to [16]. In comparison to baseline GMM based-detection in Table 4, although the MP-aware DNN could not provide better performance due to the lack of resolution within low frequencies of the used feature, it may be useful when its scores are combined with the score of another classifier.

5.2.3 Results of score combination

This subsection investigates the result of the combination of CQCC-based GMM and MP-aware DNN, compared with the combination of CQCC-based GMM and MFCC-based DNN as well as the combination of CQCC-based GMM and MGDCC-based DNN. The results are reported in Table 6.

From Table 6, the results showed that the combination of CQCC-based GMM (GMM+CQCC) and MP-aware DNN outperformed the combination of CQCC-based GMM and magnitude/phase feature-based DNN (DNN+MFCC or DNN+MGDCC). This may be because the combination of CQCC-based GMM and MP-aware DNN have a complementary nature obtained by different classifiers (that are, DNN and GMM) and features (magnitude CQT-based information, joint magnitude and phase FFT information), leading to more differences between genuine and replayed speech.

Fig. 4 shows the decision level score performance based on our proposed method. Three decision level scores including CQCC-based GMM (green line), MP-aware DNN (blue line), and the combination of CQCC-based GMM and MP-aware DNN (red line) were analyzed using 10 genuine and 10 corresponding replayed utterances from the Dev subset. As seen on the left of the figure, MP-aware DNN with different speech gives a superior boundary compared to CQCC-based GMM on genuine speech, whereas the decision performance of replayed speech is worse than CQCC-based GMM as shown on the right of the figure. It can be seen that CQCC-based GMM and MP-aware DNN have different merit-based features, which may be useful to improve the replay attack detection performance. Although the detection scores of the combination of GMM and MP-aware DNN could not give the best boundary as seen in the red line in the figure, the combined scores achieved an accuracy similar to related methods. This is because the decision level scores of CQCC-based GMM and MP-aware DNN have complementary features.

6. CONCLUSIONS AND FUTURE WORK

In this paper, an MP-aware DNN was investigated, which jointly used MFCC, and MGDCC, as a speech classification for detecting genuine and replayed speech. Also, in order take advantage of the complementary classifier-based methods to further improve the detection decision, a novel method of combining MP-aware DNN with standard replay attack detection was proposed. The result showed that the MP-aware DNN outperformed conventional DNN using only the magnitude due to the use of full information in the original signal. Although the MP-aware DNN-based method did not perform well when compared to CQCC-based GMM, the decision score was exploited to combine with the score of the CQCC-based GMM. When compared with the individual relative method, additional improvement was obtained by combining the scores of these classifiers. This paper, thus, confirmed that the MP-aware DNN is useful for replay attack detection.

Table 5: *Comparison of DNN-based detection using different features.*

Classifiers	**Features**	**Layers x Hidden nodes**	**EER (%)**	
			Dev	Eval
DNN	MFCC	**2x256**	**24.10**	**36.12**
		3x256	24.79	36.23
		2x512	24.73	36.62
		3x512	25.81	41.26
		2x1024	25.31	38.85
		3x1024	26.29	39.18
DNN	MGDCC	2x256	17.78	42.26
MP-aware DNN	MFCC+MGDCC	**2x256**	**16.76**	**35.09**

Table 6: *Performances of the proposed score combination.*

1^{st}classifier	**2^{rd}classifier**	**EER (%)**	
		Dev	Eval
GMM+CQCC	DNN+MFCC	8.36	25.56
GMM+CQCC	DNN+MGDCC	9.18	38.08
GMM+CQCC	MP-aware DNN +(MFCC+MGDCC)	**5.86**	**24.14**

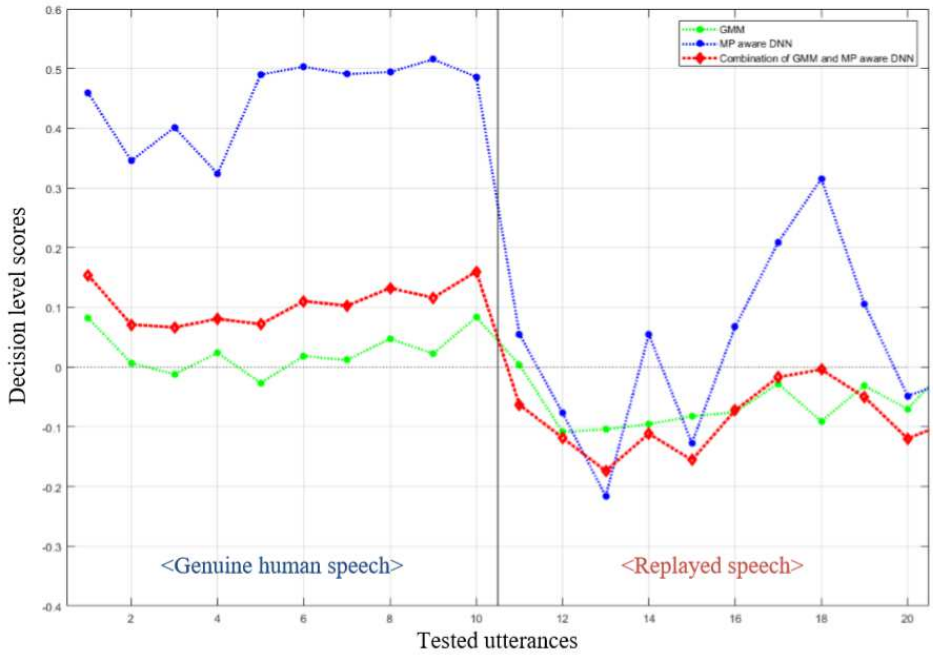

Fig.4: *Analytic illustration of decision level scores, where each replay attack detection was tested using 10 genuine and 10 corresponding replayed utterances from the Dev subset.*

In our future work, we will attempt to apply multi-task training [27–28] for DNN. Moreover, we will also make an implementation of i-vector for further performance improvement.

ACKNOWLEDGEMENT

We thank Prof. Dr. Longbiao Wang for assistance with the DNN-based speech detection program, and Miss Martha Maloi Eromine from the Institute of Research Department at RMUTI, for assistance in editing English in this manuscript.

References

[1] D. A. Reynolds and R. C. Abou Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[3] 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.

[4] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 79, no. 2, pp. 74–99, 2015.

[5] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *Proceeding of FALA workshop*, pp. 131–134, 2010.

[6] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Proceeding of 5th European Workshop on Biometrics and Identity Management*, pp. 274–285, 2011.

[7] F. Alegre, A. Janicki and N. Evans, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Proceeding of Biometrics Special Interest Group*, pp. 1–6, 2014.

[8] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5311–5327, 2016.

[9] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Proceeding of IEEE International Carnahan Conference on ecurity Technology*, pp. 1–8, 2011.

[10] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Proceeding of Asia-Pacific Information Processing Association Annual Summit and Conference*, pp. 1–5, 2014.

[11] M. Sahidullah, T. Kinnunen and C. Hanilçi, "A Comparison of features for synthetic speech detection," in *Proceeding of The International Speech Communication Association*, 2015.

[12] X. Xiao, X. Tian, S. Du, H. Xu, Chng E. S, Li. H. and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 Challenge," in *Proceeding of The International Speech Communication Association*, 2015.

[13] Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao and M. Iwahashi, "DNN-based amplitude and phase feature enhancement for noise robust speaker identification," in *Proceeding of The International Speech Communication Association*, pp. 2204–2208, 2016.

[14] D. S. Williamson, Y. Wang and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5220–5224, 2016.

[15] D. S. Williamson, Y. Wang and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[16] L. Wang, K. Phapatanaburi, Z. Oo, S. Nakagawa, M. Iwahashi and J. Dang, "Phase aware deep neural network for noise robust voice activity detection," in *Proceeding of 5th IEEE International Conference on Multimedia and Expo*, pp. 1087–1092, 2016.

[17] D. Endo, T. Hiyane, K. Atsuta and S. Kondo, "New feature for automatic speaker verification anti-spoofing: Constant Q Cepstral Coefficients," in *Proceeding of Speaker Odyssey Workshop*, pp. 249–252, 2016.

[18] K. Phapatanaburi, L. Wang, R. Sakagami, Z. Zhang, X. Li and M. Iwahashi, "Distant-talking accent recognition by combining GMM and DNN," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5109–5124, 2016.

[19] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, pp. 504–507, 2006.

[20] F. Richardson, D. Reynolds and N. Dehak, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Letters*, pp. 671–675, 2015.

[21] M. L. Seltzer, D. Yu and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7398–7402, 2015.

[22] P. Mowlaee, R. Saeidi and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *International Journal of Electronics*, vol. 81, pp. 1–19, 2016.

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

[24] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen and N. Evans, "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proceeding of International Conference on Acoustics, Speech and Signal Processing*, pp. 5395–5399, 2017.

[25] X. Tian, Z. Wu, X. Xiao, E. Chng, and H.Li, "Spoofing detection from a feature representation perspective," in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2119–2123, 2016.

[26] X. L. Zhang and J. Wu, "Deep belief networks based voice activity detection," in *Proceeding of International Conference on Acoustics, Speech and Signal Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[27] Y. Zhuang, S. Tong, M. Yin, Y. Qian and K. Yu, "Multi-task joint-learning for robust voice activity detection," in *Proceeding of International Symposium on Chinese Spoken Language Pro-*

cessing, pp. 1–5, 2013.

[28] M. J. Alam, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multi-taper MFCC features for speaker verification using I-vectors," in *Proceeding of IEEE Workshop on Automatic Speech Recognition*, pp. 547–552, 2011.

Khomdet Phapatanaburi received the B.E. degree in Electronic and Telecommunication Engineering from Rajamangala University of Technology Thanyaburi (RMUTT), Thailand in 2010, the M.E. degree in Electrical Engineering, at RMUTT in 2012, and the Dr. Eng. degree in Information Science and Control Engineering from Nagaoka University of Technology (NUT), Japan, in 2017. Since 2018, he has been a lecturer at Rajamangala University of Technology Isan Nakhonrachasrima, Nakhonrachasrima, Thailand. His main research interest is audio classification.

Prawit Buayai received the B.E. degree in Computer Engineering from Khon Kaen University (KKU), Thailand in 2013, the M.E. degree in Computer Engineering, at KKU in 2016. He is currently a Ph.D. student in Computer Engineering, University of Yamanashi, Japan. His main research is focused on the application of artificial intelligence technology to improve accessibility and productivity in agriculture.

Watcharaphon Naktong received the B.E. degree in Electrical Engineering from Rajamangala University of Technology Isan (RMUTI), Thailand in 2003, the M.E. degree in Electrical Engineering, at Rajamangala University of Technology Thanyaburi (RMUTT), Thailand in 2011. He is currently pursuing the Ph.D. degree at RMUTT, Thailand. His research interests are in the fields of Antenna Designs Ultra Wideband and UWB-MIMO.

Jakkree Srinonchat received bachelor's degree in electronic and telecommunication engineering from Rajamangala University of Technology Thanyaburi (RMUTT), Thailand, in 1995, and his Ph.D. in Electrical Engineering, major signal processing from University of Northumbria at Newcastle, UK, in 2005. He is currently a lecturer of Department of Electronics and Telecommunication Engineering, Faculty of Engineering, RMUTT, Thailand. His research is focus on the signal processing, especially FPGA Design, speech and image processing. He is currently the advisor of the Signal Processing Research Laboratory, which establishes to provide and services the new design and solution to industry.