INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017

Zhe Ji[1], *Zhi-Yi Li*[2], *Peng Li*[1], *Maobo An*[1], *Shengxiang Gao*[1], *Dan Wu*[1], *Faru Zhao*[1]

[1]National Computer network Emergency Response technical Team
Coordination Center of China(CNCERT/CC), China
[2]Big Data Innovation Center, Creditease, China

{jz,lp,amb,gsx}@cert.org.cn, zhiyili@creditease.cn, Wdcn1985@126.com, zhao@isc.org.cn

Abstract

To enhance the security and reliability of automatic speaker verification (ASV) systems, ASVspoof 2017 challenge focuses on the detection problem of known and unknown audio replay attacks. We proposed an ensemble learning classifier for CNCB team's submitted system scores, which across uses a variety of acoustic features and classifiers. An effective post-processing method is studied to improve the performance of Constant Q cepstral coefficients (CQCC) and to form a base feature set with some other classical acoustic features. We also proposed using an ensemble classifier set, which includes multiple Gaussian Mixture Model (GMM) based classifiers and two novel GMM mean supervector-Gradient Boosting Decision Tree (GSV-GBDT) and GSV-Random Forest (GSV-RF) classifiers. Experimental results have shown that the proposed ensemble learning system can provide substantially better performance than baseline. On common training condition of the challenge, Equal Error Rate (EER) of primary system on development set is 1.5%, compared to baseline 10.4%. EER of primary system (S02 in ASVspoof 2017 board) on evaluation data set are 12.3% (with only train dataset) and 10.8% (with train+dev dataset), which are also much better than baseline 30.6% and 24.8%, given by ASVSpoof 2017 organizer, with 59.7% and 56.4% relative performance improvement.

Index Terms: speaker verification, CQCC, ASVspoof, audio replay attack, ensemble learning

1. Introduction

Auto speaker verification technology makes a binary decision for accepting or rejecting a claimed identity based on a speech recording [1]. At present, it becomes more and more widely used into many forensic, civilian, and commercial applications. At the same time, the vulnerability of speaker verification systems for various types of spoofing attacks such as speaker-adapted speech synthesis, replay attacks, impersonation and voice conversion becomes more and more obvious. And reliability in the face of spoofing still remains a much great concern [2].

Generally speaking, there are mainly two strategies to protect speaker verification from spoofing. The first is to develop more robust ASV technologies and the second is to develop new spoofing countermeasures. Today, more and more research groups are paying attention to effective countermeasures of various spoofing attacks [3]. The first automatic speaker verification spoofing and countermeasures challenge in 2015 (ASVspoof 2015) successfully focuses on the countermeasures of synthesized speech and voice conversion spoofing detection with providing a common corpora and metrics [4]. Continually, ASVspoof 2017 is a second edition of this challenge. The new perspective in this challenge is audio replay attacks. The task is to determine whether a given short clip of speech audio as a 'genuine' human voice (live recording), or a 'replay' recording (fake), especially those encountered under 'unseen' conditions, for instance, containing replay environments, playback devices or talkers that might be different from those in the training data. An ideal replay detection system should be reliable to both known and unknown conditions, which are contained both in this challenge [5].

As illustrated in evaluation plan [5], despite 'ASV' being in the title, not any prior knowledge of automatic speaker verification technology is required to participate in the challenge, this task is not so much an auto speaker verification problem as a 'standalone' replay audio detection task that can be addressed as a generalized binary classification problem.

In this paper, an effective ensemble learning classifier is proposed to be a countermeasure of replay spoofing attack. This classifier ensembles a variety of base features and base classifiers. For feature extraction, we used not only classical Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual linear predictive (PLP), but also Constant Q Cepstral Coefficients (CQCC) [6][7] in baseline and an effective post-processed CQCC version. We also make an ensemble classifier set, which not only consists of baseline Gaussian Mixture Model (GMM) [8], multiple GMM based classifiers developed in auto speaker verification, like Gaussian Mixture Model-Universal Background Model (GMM-UBM) [9], GMM mean supervector-Support Vector Machine (GSV-SVM) [10] and ivector-Gaussian Probabilistic Linear Discriminant Analysis (ivector-GPLDA) [11][12], but also consists of two proposed novel classifiers: GSV-GBDT (Gradient Boosting Decision Tree) [13] and GSV-RF (Random Forest) [14]. These are all used to make up our final ensemble system.

The remainder of this paper is organized as follows. Feature extraction studies are detailed in section 2. Section 3 introduces proposed ensemble classifier set and the proposed overall ensemble learning classifier. Experimental setup and results are demonstrated in section 4. Finally, section 5 concludes the paper, while looking to the future.

2. Ensemble feature extraction

2.1. Brief overview of CQCC

Constant Q cepstral coefficients (CQCC) is first proposed in [6] for identification of musical instruments with a discrete success. It is implemented through combining the constant Q transform (CQT) with traditional cepstral analysis. In [7], a linearisation of the frequency scale of CQT is used for preserv-

http://dx.doi.org/10.21437/Interspeech.2017-1246

ing the orthogonality of the DCT basis. With this modification, new version CQCC feature can provide a variable-resolution, time-frequency representation of the spectrum which can capture more informative characteristics for spoofing detection than other traditional feature extraction used in speaker recognition. Experimental results in [7] showed that it has much better performance than traditional acoustic features for identification of synthesized and voice conversion spoofing speeches.

2.2. Post processing studies on CQCC

As shown in Table 1, we study several post-processing methods of baseline CQCC, with baseline GMM classifier on the development set of ASVspoof 2017 challenge. The experimentally results demonstrate that performing cepstral mean and variance normalization over a sliding window (WCMVN) on baseline CQCC can effectively improve its performance. In addition, only mean normalization is found to be valid as shown in row 2, while both mean and variance normalization is performing worse as the results shown with superscript*. Other method like feature warping is performing worse than baseline. WCMVN here is implemented with 301 Hamming window size and only mean normalization.

Table 1: *Comparisons of baseline CQCC and several post processing methods*

Feature	**EER(%)**	**minDCF08**
CQCC	10.4	0.996
+WCMVN	9.2	0.896
+WCMVN*	12.2	0.999
+WCMVN+PCA	7.4	0.611
+Feature warping	13.1	0.999
ensemble learning	4.6	0.471

To be further, principal component analysis (PCA) is also as one post-processing step. PCA dimension reduction projective matrix is unsupervised trained with only ASVspoof 2017 training data set, and the number of selected principal components is chosen to be 28 with the cumulative energy above 90 percent. The results show that the ensemble learning in such small base feature set, is still very effective and can achieve Equal Error Rate (EER) more than 55.8% improvement. The reason of given minDCF08 value [15] is to facilitate the observation of detection error trade-off (DET) curve trend at low false alarm rate.

2.3. Complimentary with other acoustic features

We also evaluate the replay spoofing detection performance of two traditional acoustic feature MFCC and PLP, which are both classically used in auto speaker verification. Here, the experiments are based on the GSV-SVM classifiers. The results in row 1 in Table 2 shows that with the same baseline CQCC feature, GSV-SVM performs better than baseline GMM in Table 1. It also demonstrates that although both MFCC and PLP are still performing worse for replay spoofing attack detection than CQCC, they still have good complimentary performance. In addition, it demonstrates that the ensemble learning in this feature set can also improve performance.

Table 2: *Comparisons of CQCC, MFCC and PLP*

Feature	**EER(%)**	**minDCF08**
CQCC	10.4	0.661
MFCC	27.4	0.989
PLP	37.0	0.995
ensemble learning	9.5	0.592

3. Ensemble classifier modeling

3.1. GMM baseline

As shown in given `baseline_CM.zip`, baseline GMM binary classifier trains a Gaussian mixture model on 'genuine' speech ($\mathcal{H}_{genuine}$) and another on 'spoof' speech ($\mathcal{H}_{spoof}$) by EM iterations, respectively. In the evaluation phase, the log-likelihood ratio (LLR) of a given new test segment is computed as in (1):

$$LLR = logP(D|\mathcal{H}_{genuine}) - logP(D|\mathcal{H}_{spoof}) \tag{1}$$

3.2. GMM-UBM

Different with GMM, UBM adapted GMM binary classifier firstly uses all 'genuine' and 'spoof' labeled training speeches to train a common UBM model ($\mathcal{H}_{ubm}$) by EM iterations and this UBM model is supposed to be a common label-unrelated acoustic space. Then, 'genuine' and 'spoof' models are adapted from this common UBM. In testing phase, the log-likelihood ratio (LLR) form of a new test segment is computed as in (2).

$$\begin{aligned} LLR =&(logP(D|\mathcal{H}_{genuine}) - logP(D|\mathcal{H}_{ubm}) \\ &-(logP(D|\mathcal{H}_{spoof}) - logP(D|\mathcal{H}_{ubm}) \\ =&logP(D|\mathcal{H}_{genuine}) - logP(D|\mathcal{H}_{spoof}) \end{aligned} \tag{2}$$

3.3. GSV-SVM-NAP

Based on 3.2, we also build a GSV-SVM system [10]. 'genuine' gsvs are positive inputs and 'spoof' gsvs are as negative inputs into SVM. More further, we assume that the channel information in gsvs exists in a low dimensional subspace, where channel information is defined as all the nuisance channel attribute to classification, such as replay environments, playback devices or talkers, talking content, etc. Thus, we attempt to use NAP technology to compensate the channel effect.

3.4. ivector-GPLDA

Based on 3.2, we also build an ivector-GPLDA binary classifier. We also make a similar assumption as in [11] that there exists a total variability space with no distinction between the genuine and spoof information and other channel information. Here, channel is defined as the same as in 3.3. This total variability space simultaneously captures the genuine and spoof and channel variability. Discrimination information in acoustic feature can be densely represented as a low-dimensional ivector w.

$$M = m + Tw \tag{3}$$

Simplified GPLDA modeling [12] is applied as the backend. LLR is the log likelihood ratio of probabilities of hypothesis $\mathcal{M}_{true}$ and hypothesis $\mathcal{M}_{imp}$. Here, hypothesis $\mathcal{M}_{true}$ denotes that ivector w_{mdl} and w_{seg} belong to the same class as in (4) and vice versa.

$$LLR = log\frac{P(w_{mdl}, w_{seg}|\mathcal{M}_{true})}{P(w_{mdl}, w_{seg}|\mathcal{M}_{imp})} \quad (4)$$

3.5. GSV-RF

Random Forest (RF) [14] is in itself a typical ensemble learning classifier. In training forest phrase, Tnumber of trees are built to grow while choosing $m << M$ number (M is the total feature number) of features used to calculate the best split at each node. For each tree, a training set is chosen by N times (N is the number of training examples) with replacement from the training set. For random forest, this bagging with random feature method can incorporate more diversity and reduce variances and this is useful to improve accuracy.

In our work, a GSV-RF binary classifier is proposed to append in our ensemble classifier set. It uses GMM mean supervector as input features of random forest. Output of random forest is calibrated to post probability as shown in Figure 1.

3.6. GSV-GBDT

GBDT [13] is also a typical ensemble learning classifier in itself. The idea of GBDT is boosting a set of weak learners to a strong learner and making records currently misclassified samples more important. It trains the classifier in a stage-wise fashion and it generalizes them by optimizing a differentiable loss function by iteratively choosing a function (weak hypothesis) that points in the negative gradient direction.

In our work, a GSV-GBDT binary classifier is proposed to append in our ensemble classifier set. It also uses GMM mean supervector as input features of GBDT. Output of GBDT is calibrated to post probability as shown in Figure 1.

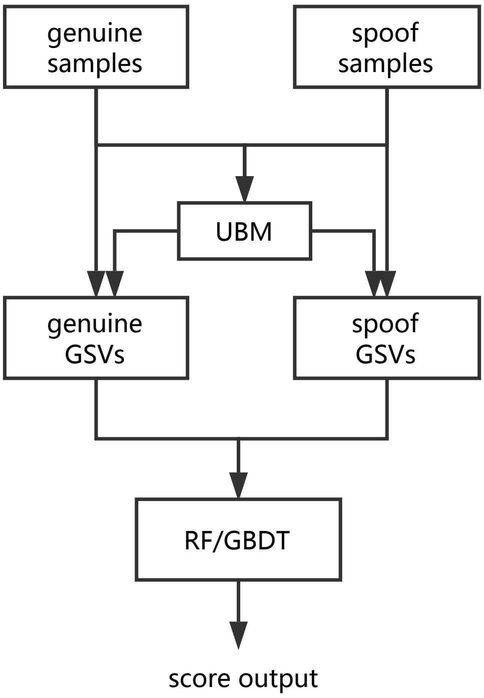

Figure 1: *Sketch of the proposed GSV-RF/GSV-GBDT binary classifier*

3.7. Our ensemble learning classifier

In our work for ASVspoof 2017, we propose an ensemble learning classifier as shown in Figure 2. We combine a variety of base feature extractions and base classifiers introduced in previous sections. All scores of base subsystems are combined with Bosaris toolkit [16] and the final output score of our ensemble learning system can be interpreted as detection log likelihood ratios.

4. Experiment setup and results

4.1. Dataset and evaluation criteria

In this work, we only focus on the common condition in ASVspoof 2017 challenge. In addition to EER, minDCF08 is also given to show comprehensive results. DET curve is also given to intuitively display the detection error trade-off for a binary classifier problem.

4.2. Experimental comparisons and analysis of base classifiers

At first, comparison of GMM and GMM-UBM with CQCC feature and several post-processing methods are demonstrated. As shown in Table 3, adapted GMM classifier is better than baseline GMM with any front end feature extraction. The results also show that the ensemble learning on this small set, is still very effective and can achieve EER more than about 57% improvement, compared to baseline GMM.

Table 3: *Comparisons of GMM and GMM-UBM with CQCC feature and several post-processing methods*

classifer	EER(%)	minDCF08
GMM	10.4	0.996
GMM_UBM	10.1	0.955
WCMVN_GMM	9.2	0.896
WCMVN_GMM_UBM	7.7	0.516
WCMVN_PCA_GMM	7.4	0.611
WCMVN_PCA_GMM_UBM	6.6	0.693
ensemble learning	4.5	0.291

It demonstrates in Table 4 the comparison of different base classifiers. It shows that GSV-SVM and ivec-PLDA can get better performance than GMM baseline and adapted GMM classifier. Although the performance of GSV-GBDT and GSV-RF is not the best, they still have good complementary performance with the other classifiers as shown with superscript[1].

Table 4: *Comparisons of classifiers with CQCC feature*

classifer	EER(%)	minDCF08
GMM[1]	10.4	0.996
GMM-UBM	10.1	0.955
GSV-SVM	10.4	0.661
GSV-SVM-NAP	9.6	0.666
ivector-GPLDA	10.2	0.482
GSV-RF[1]	9.6	0.966
GSV-GBDT[1]	13.0	0.754
ensemble learning[1]	7.1	0.563
ensemble learning (all above)	3.8	0.171

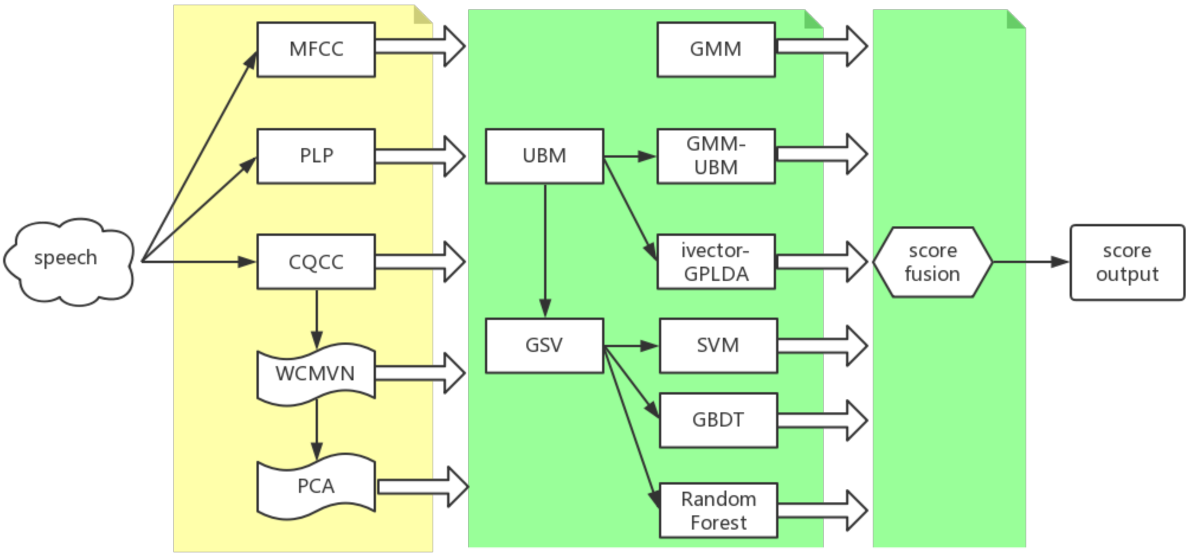

Figure 2: *Ensemble learning classifier of submitted system for ASVspoof 2017*

4.3. Experimental results of submitted systems

Table 5 shows the experimental results on the development set of our submitted primary system. S02 is denoted as the submitted primary result of our CNCB team in ASVspoof 2017 released results. EER and minDCF08 of our primary system on develop data set is 1.5% and 0.022, compared with baseline's 10.4% and 0.996, respectively.

Table 5: *Experiments on development set of baseline and submitted primary systems*

System	EER(%)	minDCF08
baseline	10.4	0.996
primary (S02)	**1.5**	**0.022**

Table 6 shows the experimental results on evaluation set of submitted primary system. EER in column 2 is performance with only train data to train models, while in column 3 with using train+development data. We can see that EERs on evaluation data set are 12.3% (with only train dataset) and 10.8% (with train+dev dataset), which are both much better than baseline 30.6% and 24.8%, given by ASVSpoof 2017 organizer, with 59.7% and 56.4% relative performance improvement.

The comparison of results, which are shown in Table 5 and Table 6, demonstrates that our proposed ensemble system may have over-fitting problem on development set. We infer that the main reason for this is the development set is too small, with only 1710 test segments in it. With more data in development set, the system can give the closer results on development set and on evaluation set.

5. Conclusions and future work

In this work, an effective ensemble learning classifier is proposed to deal with the detection problem of known and unknown audio replay attacks. This classifier uses across a variety of acoustic features and binary classifiers. The ensemble feature extraction set not only includes classical MFCC and PLP, but also CQCC and a WCMVN+PCA post-processed CQCC. And, the ensemble classifier set not only includes baseline GMM, GMM-UBM, GSV-SVM and ivector-GPLDA, but also two proposed new GSV-GBDT and GSV-RF.

On common training condition for ASVspoof 2017 evaluation, experimental results of this work have shown that the proposed ensemble learning systems can provide substantially better performance than the baseline for detection the audio replay attacks. In addition, there are also several shortcomings in this work. Firstly, the comparison of experimental results between on development set and evaluation set obviously shows that there exists an over fitting problem in training. This happens maybe because the size of development set is too small. Secondly, like phrase id and replay device, these channel information is ignored to be used for more detailed classifier modeling, such as inter dataset variability compensation, score normalization, etc. Finally, there is no time for us to try deep learning based classifier. These are all important worthy points we will focus on in the future.

Table 6: *Results on development and evaluation set of submitted primary systems*

System	EER(%) with only train	EER(%) with train+dev
baseline	30.1	24.6
primary (S02)	**12.4**	**10.8**

6. Acknowledgements

We would like to thank Dr.Md Sahidullah and the organizer of ASVspoof 2017 challenge for evaluating additional scores.

7. References

[1] Z.-Y. Li, W.-Q. Zhang, and J. Liu, "Multi-resolution time frequency feature and complementary combination for short utterance speaker recognition," *Multimedia Tools and Applications*, vol. 74, no. 3, pp. 937–953, 2015.

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

[3] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification." in *Interspeech*, 2013, pp. 925–929.

[4] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," *Training*, vol. 10, no. 15, p. 3750, 2015.

[5] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "Asvspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 1508, p. 1508, 2016.

[6] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, 2017.

[7] M. Todisco and Delgado, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 2016, pp. 249–252.

[8] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1469–1472.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[10] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[11] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[12] S. Prince, P. Li, Y. Fu, U. Mohammed, and J. Elder, "Probabilistic models for inference about identity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144–157, 2012.

[13] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 2061–2064.

[14] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1. IEEE, 1995, pp. 278–282.

[15] A. F. Martin and C. S. Greenberg, "Nist 2008 speaker recognition evaluation: Performance across telephone and room microphone channels," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[16] N. Brümmer and E. de Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.