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

Feature selection based on CQCCs for automatic speaker verification spoofing

Xianliang Wang[1], *Yanhong Xiao*[2], *Xuan Zhu*[1]

[1]Beijing Samsung Telecom R&D Center, China
[2]Beijing Institute of Technology, China

x10126.wang@samsung.com, 523078979@qq.com, xuan.zhu@samsung.com

Abstract

The ASVspoof 2017 challenge aims to assess spoofing and countermeasures attack detection accuracy for automatic speaker verification. It has been proven that constant Q cepstral coefficients(CQCCs) processes speech in different frequencies with variable resolution and performs much better than traditional features. When coupled with a Gaussian mixture model (GMM), it is an excellently effective spoofing countermeasure. The baseline CQCC+GMM system considers short-term impacts while ignoring the whole influence of channel. In the meanwhile, dimension of the feature is relatively higher than the traditional feature and usually with a higher variance. This paper explores different features for ASVspoof 2017 challenge. The mean and variance of the CQCC features of an utterance is used as the representation of the whole utterance. Feature selection method is introduced to avoid high variance and overfitting for spoofing detection. Experimental results on ASVspoof 2017 dataset show that feature selection followed by Support Vector Machine (SVM) gets an improvement compared to the baseline. It is also shown that pitch feature contributes to the performance improvement, and it obtains a relative improvement of 37.39% over the baseline CQCC+GMM system.

Index Terms: ASVspoof 2017, CQCC+GMM, feature selection, SVM

1. Introduction

Automatic speaker verification (ASV) is the technology to verification whether an utterance belongs to a given speaker [3]. It achieves great progress and has matured these years [4, 5]. Due to its low-cost, flexiblethe and rapid development, ASV is more and more widely applied in commerce. As a person authentication solution, security and reliability of ASV becomes a problem to be reckoned with. By imitating the biometric traits of client, the swindler can defraud the trust of the system. More and more researchers are concerning the vulnerability of ASV systems to different spoofing attacks [6, 7, 2], such as speech produced using text-to-speech (TTS) [8], voice conversion (VC) [9] and replay [10].

The ASV Spoofing and Countermeasures (ASVspoof) challenge provides a platform to better interlink the researchers dedicating spoofing and ASV. It was firstly organised in 2015 [11]. The ASVspoof 2015 challenge aimed to distinguish genuine speech from speech-synthesis and voice-conversion spoofing attack. The primary technical goal of ASVspoof 2017 challenge [13] is to assess spoofing attack detection accuracy in variety of unknown conditions, in particular to detect replay. And compared to speech-synthesis and voice-conversion, replay attacks are more likely to be implemented by defrauders in actual life. One may use a common electronic device to record the voice of the client, and playback the audio to the microphone.

It was stated that for ASV spoof attacts, the design of feature engineering might be more practical, and reasonable feature representation should be better than complex classifiers [2]. In[10]authors developed a spoofing attacks system using spectrum and modulation features with Support Vector Machine(SVM) model. In ASVspoof 2015 challenge, the best system [15] utilised non-conventional feature with a Gaussian mixture model (GMM). In[14], a new countermeasure was proposed for text-dependent speaker verification together with Mel-Frequency Cepstral Coefficients (MFCCs) to develop the system. MFCCs is widely used in speech signal processing as a traditional kind of feature, for example, speech recognition, speaker verification and language recognition. But the Fourier transform used in MFCCs is not ideal for spoofing detection because the frequency bins are processed similarly, and it ignores the resolution of different frequency bins.

M. Todisco proposed constant Q cepstral coefficients (CQCCs) [2]. It couples constant Q transform (CQT) with cepstral analysis, and obtains excellent performance for both known and unknown spoofing attacks. The CQT was initially proposed in music processing [16]. And it processed different frequencies with variable resolution which means higher resolution in higher frequencies while lower resolution in lower frequencies. Followed by a GMM classifier, the approach was also adopted as a baseline system for ASVspoof 2017.

This paper describes our primary system submitted to the ASVspoof 2017 challenge. Our primary system emploited feature selection based on CQCCs feature, and SVM [17] was applied as the back-end classifier.

In the submitted system, mean and variance of the CQCC feature in an utterance is concatenated as the representation of the utterance. It is thought that the channle information can be reflected by the representation. While in the CQCC+GMM, GMM is used to model the feature space. As a generative model, GMM is modelled with the short-term frames, and it depicts the several Gaussian distribution of the features. For the spoofing detection, especially for the replay task, the influence of the channel is non-ignorable. In addition, it is observed that the representation often has a high variance and the dimension is relatively high. To avoid the high variance and overfitting, feature selection is used. ReliefF algorithm [18] is adopted to select the most effective feature information. After ReliefF algorithm, Minimum Redundancy-Maximum Relevance (MRMR) [19, 20] is applied to decrease the redundant features. It is thought that the selected features reserve elements in key frequencies and remove informativeless elements. SVM is then trained to discriminate different classes.

We also explored other features for ASVspoof 2017 challenge. MFCC, pitch, emotional features [21] such as signal frame energy, zero-crossing rate, fundamental frequency were explored incorporating with CQCCs feature. Mean and variance of the features in the utterance was calculated and concatenated as the representation of the utterance.

http://dx.doi.org/10.21437/Interspeech.2017-304

The remainder of the paper is organized as follows: Details of our primary system submitted to ASVspoof 2017 challenge and other groping features are presented in Section 2. Section 3 describes the experimental corpus and the experimental results. Finally, conclusions are given in Section 4.

2. System description

The organizers provide a state-of-the-art reference implementation, CQCC+GMM to detect the replay attacks for ASVspoof 2017. The implementation is based on constant-Q transform (CQT), which is proven an effective technique in music information processing and detecting TTS and VC spoofing.

Our primary system submitted to ASVspoof 2017 challenge is based on CQCCs. Mean and variance is concatenated as the representation of the utterance. ReliefF algorithm and MRMR is then applied as feature selection method. Feature components after selection are injected as the input of SVM classifier finally.

Other features such as MFCC, pitch and emotional feature are also exploited. They are incorporated with CQCCs, and also obtain promising results.

2.1. CQCC+GMM

CQCC is based on the CQT and traditional cepstral analysis. It converts geometric space of frequency bins to a linear space by performing a linear frequency scale of the CQT. The CQCC features are variable-resolution and time-frequency representation of spectrum.

2-class GMMs for genuine and spoofed are trained. As a generative model, GMM estimates the distribution of speech from data. The parameters λ of a GMM include weight, mean and covariance matrix of Gaussion mixture component:

$$\lambda = \{\omega_i, \mu_i, \Sigma_i\} \tag{1}$$

The parameters λ are estimated with Expectation Maximization (EM) [22] algorithm.

2.2. ReliefF algorithm for feature selection

ReliefF is a improved algorithm of Relief. Relief algorithm was first proposed by Kira [23, 24]. Then Kononeil improved the Relief algorithm and proposed ReliefF algorithm. It is actually a kind of feature weighting algorithm by distinguishing between samples from close range.

The algorithm randomly chooses a sample R firstly, finds k nearest samples from the same and different classes, which is H_j and M_j, and then updates the weight of feature A by the following equation:

$$W(A) = W(A) - \sum_{j=1}^{k} diff(A, R, H_j)/(mk) + \sum_{C \in Class(R)} [\tfrac{p(C)}{1-p(Class(R))} \sum_{j=1}^{k} diff(A, R, M_j(C))]/(mk) \tag{2}$$

where $diff(A, R_1, R_2)$ represents the difference of sample R_1 and R_2 on feature A, and it is calculated by the following equation:

$$\mathrm{diff}(A, R_1, R_2) = \frac{abs(R_1[A] - R_2[A])}{\max(A) - \min(A)} \tag{3}$$

The selected feature is determined by the rank of the corresponding weight.

2.3. Minimum Redundancy-Maximum Relevance

The Minimum Redundancy-Maximum Relevance(MRMR)algorithm[19] minimizes the redundancy between features and maximizes the relevance of features and label. Mutual information is used to characterize the relevance.

The redundancy between features is calculated as following:

$$R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \tag{4}$$

The relevance between feature subset x_i and labels c is calculated through the mean of their mutual information $I(x_i, c)$:

$$D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \tag{5}$$

The final criterion combines the above two constraints as the following equation:

$$\max \Phi(\mathrm{D}, \mathrm{R}) \tag{6}$$

where

$$\Phi = \mathrm{D} - \mathrm{R} \tag{7}$$

2.4. Support Vector Machines

Different from GMM, SVM is a discriminative model. It is a popular technique for discriminative classification. It is a classifier that find a maximum margin. The best separator of a SVM is defined by a kernel function as follows:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i t_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b \tag{8}$$

where N is the number of support vectors, α_i and b are the SVM parameters during the training step, and t_i is the label of the support vector $\boldsymbol{x}_i$. The value of the label is either 1 or -1, depending upon whether the corresponding support vector belongs to class 1 or -1.

The kernel function $K(\cdot, \cdot)$ can be expressed as

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \phi(\boldsymbol{x})' \phi(\boldsymbol{x}_i) \tag{9}$$

where $\phi(\boldsymbol{x})$ is a mapping from the input space to a possibly infinite dimensional SVM expansion space.

3. Experiment

This section presents the data and setup used in the experiments firstly. Experimental results of our primary system and exploring experiments on development data and evaluation data are given.

3.1. Data and evaluation metric

Data used in our experiments is provided by the organizers of ASVspoof 2017 challenge including training, development and evaluation subsets. The detection of ASVspoof 2017 is a 2-class problem for genuine and spoofed. The number of genuine and spoofed utterances in the training, development and evaluation subset is given in Table 1.

The common training set containing 1,508 genuine and 1,508 spoof files was used in our system, which was a required

submitted system. If not specified, the following systems were trained with only the common training set.

Equal error rate (EER) is used as the performance measures in the ASVspoof 2017 challenge, which is also the evaluation metric in our experiments.

Table 1: *Number of genuine and spoofed utterances in the training and development subset*

Subset	#Training	#Development	#Evaluation
genuine	1,508	760	1,298
spoofed	1,508	950	12,008

3.2. Experimental setup and results

Some different kinds of features were explored on the development set. We explored the effect of MFCC, emotional features and pitch. In Table 2, results of MFCC+GMM, EMOTION+SVM, CQCC180+SVM, FEAT564+SVM and CQCC180+Pitch+SVM system are presented.

Table 2: *Performance of MFCC+GMM and EMOTION+SVM system on development set*

system	EER
CQCC+GMM	11.02
MFCC+GMM	16.60
EMOTION+SVM	14.18
CQCC180+SVM	10.14
FEAT564+SVM	10.60
CQCC180+Pitch+SVM	**9.27**

In MFCC+GMM, dimension of MFCC used in the experiment was 60, including 19 Mel-Frequency Cepstral Coefficients and the log-energy, the first and the second derivatives. The emotional features in EMOTION+SVM included 384 dimension, which were extracted using the toolkit openSMILE [25]. Details of the emotional feature are as Table 3. The CQCC180+SVM calculated the CQCC mean and variance of the frames in every utterance without feature selection, and it resulted in 180-dimensional feature for every utterance. SVM was used to train the back-end SVM classifier. In the experiment, libsvm with radial basis function (RBF) was applied [26]. We then concatenated the 180-dimensional feature and 384-dimensional emotional feature and SVM was used to model the concatenated features in FEAT564+SVM system.

Table 3: *low-level descriptors of emotional feature*

descriptors	description
pcm_RMSenergy	Root-mean-square energy
MFCC	12-dimensional MFCCs
pcm_zcr	Zero-crossing rate
voiceProb	Voice probablity
F0	Fundamental frequency

From the results in Table 2, it can be seen that the CQCC feature performs much better than the traditional MFCC and emotional feature. Even though emotional feature performs not so good as CQCCs, it is thought some attributes are informative to spoofing attacks and are promising when combing with attributes selection or other features. We will further explore the emotional feature in the future. Mean and variance presentation followed by SVM got a relative improvement of 8.00%. Since the mean and variance presentation can be seen as characteristic reflection of the whole utterance, specially the channel information, it is critical for discriminating replay attacks. The pitch brought a relatively improvement of 8.58%. Figure 1 showed the differences in pitch between genuine and spoofed utterances which were chosen from the same person and the same sentence, five statistic features from the non-zero pitch of each utterance including maximum, minimum, mean, variance and range were calculated combing with CQCC mean and variance as 185 dimensional features to detect spoof attack.

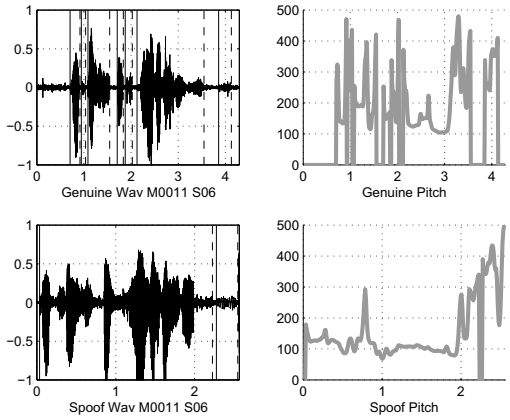

Figure 1: *Pitch of genuine and spoofed utterance*

In the following, experiments of feature selection are presented. In Figure 2, performance with rank of feature selection is depicted. According to the results, 100 is selected as the rank of feature selection in the following results.

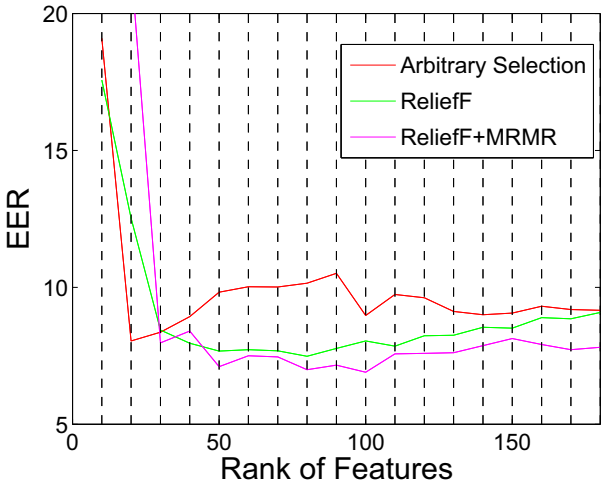

Figure 2: *Performance with rank of feature selection*

From Table 4, it is seen that feature selection improves the performance. Taking the CQCC180+ReliefF+MRMR+SVM as an example, it obtained a relative performance improvement of 6.80% compared with CQCC180+SVM system and 14.25% compared with the baseline CQCC+GMM system.

Table 4: *Performance of feature selection method on development set*

system	EER
CQCC180+ReliefF+SVM	9.47
CQCC180+ReliefF+MRMR+SVM	9.45
FEAT564+ReliefF+SVM	10.59
FEAT564+ReliefF+MRMR+SVM	10.20

Pitch is also a productive factor to the replay attack feature. After ReliefF, we got ranks of 185-dimensional features , all the statistic features of pitch were in Top100 and the rank of pitch mean and variance is sixth and eighteenth. Results in Table 5 showed the pitch achieved a 26.98% relative improvement, and combining with CQCC and feature selection, it obtained a relative improvement of about 37.39% totally compared with CQCC+GMM system. The ReliefF got a relative improvement of 13.27% and MRMR improved by 14.18%. Performance of feature selection method and pitch is presented in Table 4 and 5 respectively. DET curve of the system is given in Figure 3.

Table 5: *Performance of adding pitch feature on development set*

system	EER
CQCC180+Pitch+ReliefF+SVM	8.04
CQCC180+Pitch+ReliefF+MRMR+SVM	**6.90**

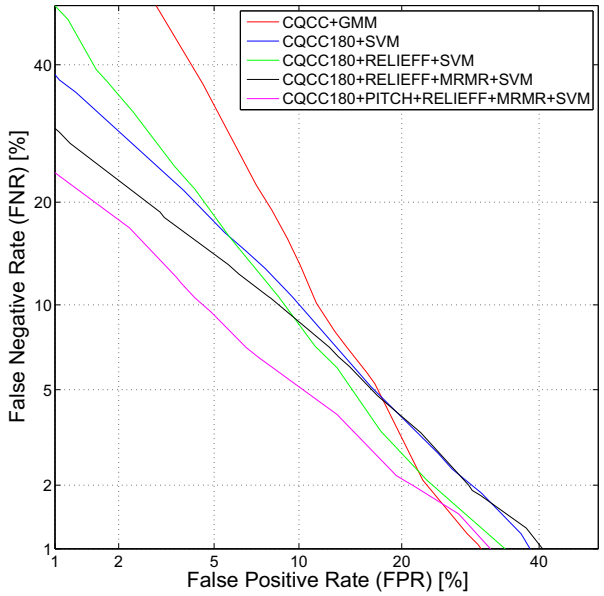

Figure 3: *DET curves of the systems*

Results of the baseline system (CQCC+GMM), primary system CQCC180+ReliefF+MRMR+SVM and CQCC180+Pitch+ReliefF+MRMR+SVM system on the updated evaluation set are given in Table 6. It is shown that the CQCC180+ReliefF+MRMR+SVM system got a relatively improvement of 3.62% compared to the CQCC+GMM system and pitch with feature selection brings a relatively improvement of 13.03%.

Table 6: *Performance of the CQCC+GMM, CQCC180+ReliefF+MRMR+SVM and CQCC180+Pitch+ReliefF+MRMR+SVM system on evaluation set*

system	EER
CQCC+GMM	28.48
CQCC180+ReliefF+MRMR+SVM	27.45
CQCC180+Pitch+ReliefF+MRMR+SVM	24.77

4. Conclusions

The ASV spoofing aims to determine whether a speech audio is a genuine human voice or a spoofing. With a growing researchers working to develop spoofing countermeasures, the ASVspoof challenge provides a platform to better interlink the researchers dedicating spoofing and ASV.

This paper presents the primary system description for ASVspoof 2017 challenge. Our primary submitted system is based on CQCC feature, and feature selection method is applied to purifying the feature representation. The global representation of an utterance is proven effective to spoofing attacks. The introduction of new feature factor such as pitch is beneficial to the system performance. Experimental results show pitch factor and feature selection method significantly improves the performance.

In the future, more efforts will be taken to ASV spoofing attacks detection. More new features for ASV spoofing attacks will be explored and we will try feature attributes selection in emotional feature combing with CQCCs or other new features.

5. Acknowledgements

The authors would like to thank the organizers of the ASVspoof 2017 challenge.

6. References

[1] M. Todisco, "Articulation rate filtering of cqcc features for automatic speaker verification," pp. 3628–3632, 2016.

[2] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Odyssey 2016 - The Speaker and Language Recognition Workshop*, 2016.

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

[4] W. Rao, M. W. Mak, and K. A. Lee, "Normalization of total variability matrix for i-vector/plda speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4180–4184.

[5] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," *Computer Science*, 2015.

[6] N. W. D. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *INTERSPEECH 2013, Conference of the International Speech Communication Association, August 25-29, 2013, Lyon, France*, 2013.

[7] 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.

[8] P. L. D. Leon, M. Pucher, J. Yamagishi, and I. Hernaez, "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.

[9] B. L. Pellom and J. H. L. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference*, 1999, pp. 837–840.

[10] J. Villalba and E. Lleida, *Detecting Replay Attacks from Far-Field Recordings on Speaker Verification Systems*. Springer Berlin Heidelberg, 2011.

[11] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc, and M. S. A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH*, 2015.

[12] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Biometrics Special Interest Group*, 2015, pp. 1–6.

[13] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "Asvspoof: the automatic speaker verification spoofing and countermeasures challenge," in *IEEE Journal of Selected Topics in Signal Processing Special Issue on Spoofing and Countermeasures for Automatic Speaker Verification*, 2017.

[14] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Asia-Pacific Signal and Information Processing Association, 2014 Summit and Conference*, 2015, pp. 35–45.

[15] T. B. Patel and H. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *The Conference of International Speech Communication Association*, 2015.

[16] J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *Journal of the Acoustical Society of America*, vol. 105, no. 3, p. 1933, 1999.

[17] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[18] I. Kononenko, *Estimating attributes: Analysis and extensions of RELIEF*. Springer Berlin Heidelberg, 1994.

[19] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 8, pp. 1226–38, 2005.

[20] A. Unler, A. Murat, and R. B. Chinnam, "mr 2 pso : A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification," *Information Sciences*, vol. 181, no. 20, pp. 4625–4641, 2011.

[21] B. Schuller, S. Steidl, A. Batliner, and F. Jurcicek, "The interspeech 2009 emotion challenge – results and lessons learnt," *Interspeech*, pp. 312–315, 2009.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[23] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm," in *Tenth National Conference on Artificial Intelligence*, 1992, pp. 129–134.

[24] ——, "A practical approach to feature selection," in *International Workshop on Machine Learning*, 1992, pp. 249–256.

[25] F. Eyben and B. Schuller, "opensmile:): the munich open-source large-scale multimedia feature extractor," *Acm Sigmultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.

[26] C. C. Chang and C. J. Lin, "Libsvm: A library for support vector machines," *Acm Transactions on Intelligent Systems & Technology*, vol. 2, no. 3, article 27, p. 27, 2007.