Siamese Convolutional Neural Network Using Gaussian Probability Feature for Spoofing Speech Detection

Zhenchun Lei[1], *Yingen Yang*[1], *Changhong Liu*[1], *Jihua Ye*[1]

[1] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China
zhenchun.lei@hotmail.com, yyg1999@sina.com, liuch@jxnu.edu.cn, yjhwcl@163.com

Abstract

The security and reliability of automatic speaker verification systems can be threatened by different types of spoofing attacks using speech synthetic, voice conversion, or replay. The 2-class Gaussian Mixture Model classifier for genuine and spoofed speech is usually used as the baseline in the ASVspoof challenge, which is designed to develop the generalized countermeasures with potential to detect varying and unforeseen spoofing attacks. In the scoring phase, the GMM accumulates the scores on all frames in a test speech independently, and does not consider the relationship between adjacent frames. We propose the 1-D Convolutional Neural Network whose input is the log-probabilities of the speech frames on the GMM components. The new model considers not only the score distribution of GMM components, but also the local relationship of frames. And the pooling is used to extract the speech global character. The Siamese CNN is also proposed, which is based on two GMMs trained on genuine and spoofed speech respectively. Experiments on the ASVspoof 2019 challenge logical and physical access scenarios show that the proposed model can improve performance greatly compared with the baseline systems.

Index Terms: anti-spoofing, spoofing speech detection, convolutional neural network, gaussian mixture model

1. Introduction

Automatic Speaker verification (ASV) aims to automatically confirm the identity of the speaker by given a speech segment [1]. The recent advances in speech technologies have posed a great threat to the ASV system with various spoofing attacks. There are four well-known attacks that present a serious threat to ASV systems, namely, mimicry [2], text-to-speech (TTS) [3], voice conversion (VC) [4], replay [5]. To counteract these spoofed attacks, countermeasures (CM) have been developed to detect spoofed attacks before speaker verification.

A multitude of different anti-spoofing technologies have been put forward to improve detection performance. Most of the efforts are focus on designing discriminative features [6]. Recent work has also shown that high-frequency sub-bands in the acoustic signal contain more evidence, and become popular solutions. For example, Constant Q cepstral coefficients (CQCCs) [7], which use the constant Q transform (CQT) instead of the short-time Fourier transform (STFT) to process speech signals, perform better than common Mel-Frequency Cepstral Coefficients (MFCCs) [8]. Linear Frequency Cepstral Coefficients (LFCCs) [6], Inverse Mel Frequency Cepstral Coefficients (IMFCCs) [9], Rectangular Filter Cepstral Coefficients (RFCCs) [10], Linear Prediction Cepstral Coefficients (LPCCs) [11], Sub-band Spectral Centroid Magnitude Coefficients (SCMCs) [12], and Scattering Cepstral Coefficients (SCC) [13] have been shown to be effective front-ends for spoofing detection.

In the aspect of classifiers, the classical Gaussian Mixture Model (GMM) [14] is usually used as the baseline system. With the development of deep learning technology, more and more neural network models are applied to spoofing speech detection. Galina [15] employed Light Convolutional Neural Networks (LCNN) with max filter map activation function, which get the best performance in ASVspoof 2017 challenge. Alejandro [16] proposed a hybrid LCNN plus RNN architecture which combines the ability of the LCNNs for extracting discriminative features at frame level with the capacity of gated recurrent unit (GRU) based RNNs for learning long-term dependencies of the subsequent deep features. Moustafa [17] proposed deep residual neural networks for audio spoofing detection, which process MFCC, CQCC and spectrogram input features, respectively. Cheng-I [18] proposed attentive filtering network, which is composed of an attention-based filtering mechanism that enhances feature representation in both the frequency and time domains, and a ResNet-based classifier. Neural network classifiers are reported to give better performance than GMM.

Siamese networks are first proposed by Bromley et al. [19] for signature verification. Siamese CNN contains more than two branches of CNNs which they are often identical. Each branch includes series of convolutional, ReLU, pooling, and fully connected layers [20]. These multiple branches of CNNs are trained simultaneously and they create the same dimension feature vectors. There are many Siamese models such as Siamese, Pseudo-Siamese, and 2-channel [21]. Chen and Salman [22] proposed a regularized Siamese deep network to extract speaker-specific information from MFCCs for a speaker recognition task. Yichi Zhang [23] proposed Siamese style CNNs for sound search by vocal imitation. Kaavya [24] proposed a deep Siamese network to identify pairs of genuine speech samples and pairs of replayed speech samples as being similar and mixed pairs of genuine and replayed speech to be identified as dissimilar.

In the classical GMM, the score is accumulated on all feature frames independently, and each Gaussian component's contribution information is discarded. The relationship between adjacent frames is also been ignored along the time axis. To address these issues, we proposed 1-D CNN and Siamese CNN using gaussian probability feature for spoofing speech detection.

The remainder of the paper is organized as follows. In section 2, extracting of Gaussian probability feature is presented. Section 3 descripts the CNN and Siamese CNN for spoofing detection. In section 4, the experimental setup, results and discussion are presented. Finally, section 5 provides the conclusions.

2. Gaussian Probability Feature

2.1. Gaussian Mixture Model

The GMM provides an effective way to describe the speech characters, and one of its powerful attributes is the capability to form smooth approximations to arbitrarily shaped densities. For a d-dimensional feature vector, x, the mixture density used for the likelihood function has the following form:

$$\mathrm{p(x)} = \sum_{i=1}^{M} w_i p_i(x) \quad (1)$$

The density is a weighted linear combination of M unimodal Gaussian densities, $p_i(x)$,each parameterized by a mean $D \times 1$vector, μ_i, and a $D \times D$ covariance matrix, Σ_i:

$$\mathrm{p}_i(\mathrm{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\} \quad (2)$$

The parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}$ of GMM are estimated with Expectation Maximization (EM) algorithm.

The baseline system includes two GMMs: one for genuine speech and one for spoof speech. For a given test speech utterance $X = \{x_1, x_2, \dots, x_N\}$, the log-likelihood ratio is used to make the human/spoof decision, and the log-likelihood ratio is defined in the flowing form:

$$\mathrm{score}_{baseline} = \log p(X|\lambda_h) - \log p(X|\lambda_s) \quad (3)$$

where λ_h and λ_s are the GMMs for human and spoof speech respectively.

2.2. Gaussian Probability Feature

For a speech feature sequence, the GMM accumulates the scores on all frames independently, and does not consider the contribution of every Gaussian component to the final score. Moreover, the relationship between adjacent frames is also been ignored. So, we want to model the score distribution on every GMM component, and propose the Gaussian probability feature.

For a raw frame feature x_i (CQCC or LFCC in our experiments), the size of new feature f_i is the order of GMM and the component f_{ij} is:

$$\mathrm{f}_{ij} = \log(w_j \cdot p_j(x_i)) \quad (4)$$

After that, the mean and standard deviation of the training data are calculated and used for mean and variance normalization for each utterance.

3. Siamese Convolutional Neural Network

3.1. 1-D Convolutional Neural Network

With the development of deep learning technology, Convolutional Neural Networks based models have been popularly applied in various machine learning tasks. CNN utilize layers with convolving filters that are applied to local features. We propose the 1-D CNN whose input is the log-probabilities of the speech frames on the GMM components. The new model considers not only the frame scores on GMM, but also the local relationship between frames.

We build our 1-D CNN model upon that of which is proposed for sentence classification [25]. Figure 1 shows the architecture of CNN model.

The GMM is trained on the whole training dataset without labels. The convolutional layer takes the log-probabilities as input features. A convolution operation involves a filter which is applied to a window of features to produce a new feature. Then a max-overtime pooling operation is applied over the feature map and the maximum value is taken as the feature corresponding to this particular filter. The idea is to capture the most important feature—one with the highest value—for each feature map. This pooling scheme naturally deals with variable speech lengths. The model uses multiple filters (with varying window sizes) to obtain multiple features. These concatenated features form the penultimate layer and are passed to a fully connected layer with dropout and softmax output. Final the output of model is the probability distribution over genuine and spoofed speech labels. In our experiments, sizes of the filter windows are 3, 4, 5, 6, 7, and each filter has 512 feature maps. The dropout rate in fully connected layer is 0.5.

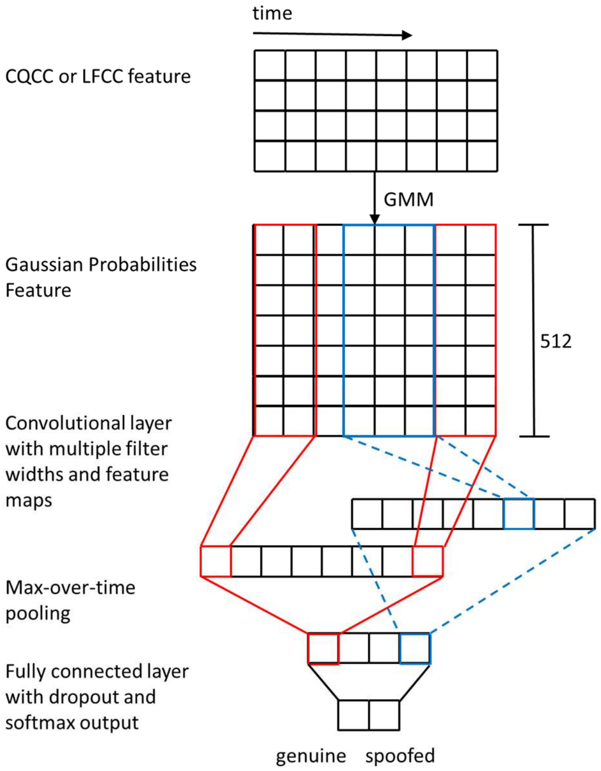
time
CQCC or LFCC feature
GMM
Gaussian Probabilities Feature
512
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output
genuine
spoofed

Figure 1: *CNN Model architecture for spoofing detection.*

3.2. Siamese Convolutional Neural Network

We also propose a Siamese CNN for spoofing speech detection, which is based on two GMMs trained on genuine and spoofed speech respectively. The proposed network architecture is depicted in Figure 2.

The Siamese CNN contains two identical CNNs, each of which has the same architecture in previous section except for the fully connected layer. The input of convolutional layer is log-probabilities calculated separately by two GMMs in the baseline system. The branches of CNNs are trained simultaneously on training dataset and they create two same dimension embedding vectors. Then we concatenate these two vectors and input it to the fully connected layer with dropout and softmax output.

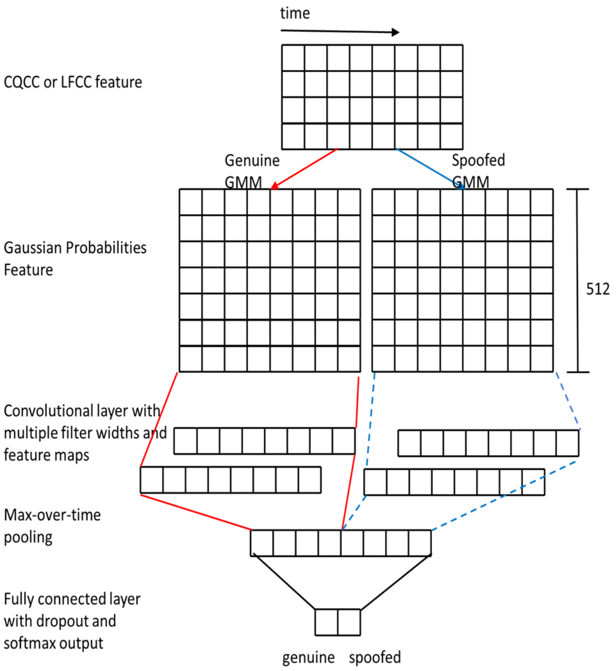
time
CQCC or LFCC feature
Genuine GMM
Spoofed GMM
Gaussian Probabilities Feature
512
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output
genuine
spoofed

Figure 2: *Siamese CNN Model architecture for spoofing detection.*

4. Experiments

4.1. Setup

The experiments were run on the ASVspoof 2019 [26] database consisting of bona fide and spoofed speech signals which are derived from the VCTK base corpus. The database encompasses two partitions for the assessment of logical access (LA) and physical access (PA) scenarios. The LA scenario involves spoofing attacks that are injected directly into the ASV system. Attacks in the LA scenario are generated using the latest text-to-speech synthesis (TTS) and voice conversion (VC) technologies. For the PA scenario, speech data is assumed to be captured by a microphone in a physical, reverberant space. ASVspoof 2019 corpus is divided into three no speakers overlap subsets: training, development and evaluation. And three data subsets are all split into two parts, namely the bona fide and the spoofed speech. More detailed description of these subsets can be found in [26]. Performance is measured in terms of minimum t-DCF and EER.

In the experiments, we used constant Q cepstral coefficients (CQCC) and linear frequency cesptral coefficients (LFCC) as acoustic features for anti-spoofing. The CQCC is extracted with the CQT analysis of speech, and has been shown to perform competitively better than other features in speaker spoofing detection. LFCC is similar to MFCC except for the use of a linear-scaled in place of a Mel-scaled filterbank, thereby giving a constant spectral resolution. LFCC has also been applied to speech recognition, speaker recognition, and spoofing detection. The feature extractors are matlab implementation of spoofing detection baseline system provided by the organizers. The feature extractors use the default configuration.

The baseline is two separate GMMs system, which are trained using maximum-likelihood criteria from genuine and spoofed speech-data respectively. We train GMMs with 512 mixtures and 30 EM iterations using the MSR Identity Toolbox [27] implementation. The log-likelihood ratio of each tested speech sample from the genuine model and spoofed model is taken as the final score during evaluation.

We implemente our neural network models using PyTorch and train models using a machine with a GTX 1080 Ti GPU. Cross-entropy loss is adopted as the loss criterion and Adam optimizer with learning rate of 0.0001 is used during the training process. The batch size is set to 32 in all experiments.

4.2. Results on ASVspoof 2019 LA scenario

The LA scenario contains bona fide speech and spoofed speech data generated using 17 different TTS and VC systems. There are 6 algorithms (3 speech synthesis algorithms, 1 TTS system implementation, and 2 voice conversion) included in the training and development data. To obtain more generalized systems under mismatched conditions, the evaluation contains 14 kinds of spoofed speech generated by unseen spoofing algorithms. Table 1 and 2 show the results on the ASVspoof 2019 LA scenario obtained by the baseline systems and our Siamese CNN systems using CQCC and LFCC respectively.

Table 1: *Results on ASVspoof 2019 Logical Access in terms of min-tDCF and EER (%) using CQCC*

Model	dev		eval	
	EER	min-tDCF	EER	min-tDCF
GMM	0.237	0.00665	8.97	0.214
CNN	0.126	0.00355	9.61	0.217
Siamese CNN	0.157	**0.00287**	**8.75**	**0.211**

Table 2: *Results on ASVspoof 2019 Logical Access in terms of min-tDCF and EER (%) using LFCC*

Model	dev		eval	
	EER	min-tDCF	EER	min-tDCF
GMM	3.77	0.103	7.59	0.208
CNN	1.066	0.279	4.28	0.122
Siamese CNN	**0.710**	**0.019**	**3.79**	**0.093**

The proposed Siamese CNN system outperforms the baseline systems on the development and evaluation datasets obviously. Specifically, the LFCC + Siamese CNN system improves the baseline system min-tDCF and EER by 55.29% and 50.06% on the evaluation dataset, respectively. Both CNN and Siamese CNN can get little performance improvement using CQCC feature compared with LFCC + GMM system. Maybe our new model is more compatible with LFCC feature.

4.3. Results on ASVspoof 2019 PA scenario

The physical access condition considers spoofing attacks that are performed at the sensor level. Spoofing attacks in this scenario are therefore referred to as replay attacks, whereby a recording of a bona fide access attempt is first captured, presumably surreptitiously, before being replayed to the ASV microphone. Table 3 and 4 show the results on the ASVspoof 2019 PA scenario obtained by the baseline systems and our Siamese CNN systems using CQCC and LFCC respectively.

Table 3: *Results on ASVspoof 2019 Physical Access in terms of min-tDCF and EER (%) using CQCC*

Model	dev		eval	
	EER	min-tDCF	EER	min-tDCF
GMM	9.72	0.184	11.34	0.254
CNN	10.52	0.247	10.86	0.267
Siamese CNN	9.89	**0.218**	**10.08**	**0.245**

Table 4: *Results on ASVspoof 2019 Physical Access in terms of min-tDCF and EER (%) using LFCC*

Model	dev		eval	
	EER	min-tDCF	EER	min-tDCF
GMM	11.22	0.236	12.90	0.287
CNN	8.74	0.202	9.48	0.241
Siamese CNN	**7.35**	**0.167**	**7.98**	**0.195**

Similar performance improvement can be obtained on the PA dataset. Siamese CNN can get the best performance, and CNN get little worse results. And Siamese CNN get more performance improvement compared with the baseline system using LFCC feature. The LFCC + Siamese CNN system achieves a relative 38.14% and 32.06% better performance than LFCC + GMM on the evaluation set in min-TDCF and EER, respectively.

5. Conclusions

In this paper, we proposed CNN and Siamese CNN models using Gaussian probability feature for spoofing speech detection. The classical GMM accumulates the scores on all frames independently, and does not consider the contribution of every Gaussian component to the final score. And the relationship between adjacent frames is also been ignored. For an utterance, the Gaussian probability feature includes the score distribution on each GMM component. We propose the 1-D CNN model which considers not only the frame scores on GMM, but also the local relationship between frames. We also propose a Siamese CNN for spoofing speech detection, which is based on two GMMs trained on genuine and spoofed speech respectively. The experimental results on the ASVspoof 2019 database show that the proposed Siamese CNN can improve performance greatly.

For future work, it is essential to explore new neural network architecture to model Gaussian probability feature. ResNet has good performance in many machine learning tasks, and we will consider the similar architecture combing with the new feature. Many studies have showed that the combined CNN and LSTM model produced a more robust model. High performance of CNN-LSTM model is due combining capability of CNN to capture short-term feature relations and LSTM to capture longer-term temporal feature relations. We will explore the CNN-LSTM model in spoof speech detection. On the other hand, these new models will also be applied to speaker recognition in future.

6. Acknowledgements

This work is supported by National Natural Science Foundation of P.R.China (61365004, 61662030), and by Educational Commission of Jiangxi Province of P.R.China (GJJ170205).

7. References

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, “Spoofing and counter-measures for speaker verification: A survey,” Speech Communication, Vol. 66, no. 0, pp. 130– 153, 2015.

[2] Hautamäki RS et al, “Automatic versus human speaker verification: the case of voice mimicry,” Speech Commun, 72:13–31, 2015.

[3] M. Sahidullah, T. Kinnunen, and C. Hanilci, “A comparison of features for synthetic speech detection," in Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc., Dresden, Germany, 2015, pp. 2087-2091.

[4] Toda T, Black AW, Tokuda K, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE Trans Audio, Speech, Lang Process 15(8):2222–2235, 2007.

[5] J. Galka, M. Grzywacz, R. Samborski, “Playback attack detection for text-dependent speaker verification over telephone channels,” Speech Communication, vol 67, pp.143–153, 2015.

[6] M. Sahidullah, T. Kinnunen, C. Hanil¸ci, “A comparison of features for synthetic speech detection,” in: Proc. INTERSPEECH, pp.2087–2091, 2015.

[7] M. Todisco, H. Delgado, and N. Evans, “Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," Computer Speech Language, vol. 45, pp. 516-535, 2017.

[8] S. B. Davis and P. Mermelstein, “Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.

[9] S. Chakroborty, A. Roy, and G. Saha, “Improved closed set textindependent speaker verification by combining MFCC with evidence from flipped filter banks,” International Journal of Signal Processing, vol. 4, no. 2, pp. 114–122, 2007.

[10] T. hasan, S. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. Hansen, “CRSS systems for 2012 NIST speaker recognition evaluation,” in ICASSP 2013 – International Conference on Acoustics, Speech and Signal Processing, 2013.

[11] S. Furui, “Cepstral analysis technique for automatic speaker verification,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 2, pp. 254–272, 1981.

[12] E. A. P. N. Le, J. Epps, V. Sethu, , and E. Choi, “Investigation of spectral centroid features for cognitive load classification,” Speech Communication, vol. 53, no. 4, pp. 540–551, 2011.

[13] K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li, “Front-End for Anti-Spoofing Countermeasures in Speaker Verification: Scattering Spectral Decomposition,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no.4,pp.632-643, June 2017

[14] Marcin Withowski, Stanislaw Kacprasko, Piotr Zelasko, Konrad Kowlczyk, and Jakub Galka, “Audio replay attack detection using high-frequency features,” in Interspeech, 2017, pp. 27–31.

[15] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, “Audio Replay Attack Detection with Deep Learning Frameworks,” Interspeech, pp. 82–86, 2017.

[16] Alejandro Gomez-Alanis1, Antonio M. Peinado1, Jose A. Gonzalez, and Angel M. Gomez, “A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection”, in interspeech, pp. 1068-1072, 2019

[17] Moustafa Alzantot1, Ziqi Wang, Mani B. Srivastava, “Deep Residual Neural Networks for Audio Spoofing Detection,” in interspeech, pp.1078-1082, 2019

[18] Cheng-I Lai, Alberto Abad, Korin Richmond, Junichi Yamagishi, Najim Dehak, and Simon King, “Attentive Filtering Networks for Audio Replay Attack Detection,” in ICASSP, pp. 6316-6320, 2019

[19] J. Bromley, I. Guyon, and R. Shah, “Signature Verification Using a ‘Siamese’ Time Delay Neural Network,” Int. J.Pattern Recognit. Artif. Intell., vol. 7, no. 4, pp. 669–688, 1993.

[20] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, 2016.

[21] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4353-4361, 2015.

[22] K. Chen and A. Salman, "Extracting speaker-specific information with a regularized siamese deep network", Proc. Adv. Neural Inf. Process. Syst., pp. 298-306, 2011.

[23] Y. Zhang, B. Pardo and Z. Duan, "Siamese Style Convolutional Neural Networks for Sound Search by Vocal Imitation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 2, pp. 429-441, Feb. 2019.

[24] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric," Interspeech, pp. 671–675, 2018.

[25] Yoon Kim, "Convolutional Neural Networks for Sentence Classification," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751. ACL, Doha, Qatar, 2014.

[26] ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. [Online]. http://www.asvspoof.org/asvspoof2019/asvspoof2019evaluation\plan.pdf, 2019

[27] S.O.Sadjadi et al., "MSR Identity Toolbox v1.0: A matlab toolbox for speaker recognition research," *Speech and Language Processing Technical Committee Newsletter*, 2013.