STC ASVspoof 2017 systems descriptions

Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, Vadim Shchemelinin

STC Ltd., St.Petersburg, Russia
www.speechpro.com

1 Introduction

We submitted evaluation scores for six different spoofing detection systems: three for common conditions and three for flexible. Four proposed systems are the combination of several subsystems of two types (Picture 1). The final score of these systems is the result of all subsystems scores fusion.
Subsystems of the first type use features, extracted from the input signal for further high level features extraction based on neural network or i-vector approaches, then high level features are used by the classifier to make the decision if the signal belongs to the genuine speech class or not. Subsystems of the second type use neural networks as end-to-end module to get the decision directly from the extracted features. Concrete details of all subsystems are described in Section 2.
The essential note common for all systems that was found during the preliminary experiments is that mean variance normalization can be highly important for stable training. Because of that it is used for all types of acoustic features in the Front-End of proposed systems.
To train systems for common conditions of ASVspoof Challenge 2017 we used only the train part of the Challenge. The dev part was used for performance validation and weights adjustment in system fusion. For flexible conditions we pooled train part with additionally collected databases, described in Section 3.

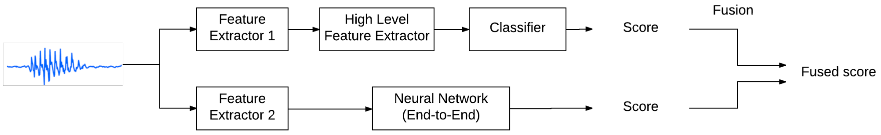

Fig. 1: CNN architecture for single-photo liveness detection

2 Subsystems description

2.1 CQT-CNN-GMM-SW subsystem

This subsystem uses the log power magnitude spectrum of constant Q transform [5] as informative time-frequency representation. The quantized grayscale images are extracted from the spectrum with 200 size sliding window and 20 length step. We apply mean variance normalization to these images and use them as an input for high level feature extractor. Convolutional Neural Network with inception modules ([2]) is used further for high level feature extraction. This high level features are used next as an input for 2-class GMM classifier (1 GMM for genuine speech and 1 for spoofed speech) trained on the training part of the ASvspoof 2017 database. We use 1 component models, trained with EM (Expectation Maximization) algorithm.
The score for an input signal is computed as the loglikelihood ratio $\Lambda(X) = \log L(X|\Theta_n) - \log L(X|\Theta_s)$, where X is a sequence of test utterance feature vectors, L denotes the likelihood function, and Θ_n and Θ_s represent the GMMs for natural and spoofed speech respectively.

2.2 CQT-CNN-GMM subsystem

In this subsystem we also use constant Q transform, but we don't use sliding window approach and extract one image of width 400 from each spectrogram. The high level feature extractor is based on Light CNN [7].

2.3 DWT-CNN-GMM-SW subsystem

This subsystem is similar to the CQT-CNN-GMM system and consists of the CNN-based high level features extractor and 2-class GMM classifier. In the feature extraction module we use Discrete wavelet Transform spectrum, obtained by Daubechies wavelets db4. To obtain final image representation we apply sliding window with 200 window size and 33.3 length step. High level feature extractor is based on the Light CNN implementation.

2.4 FFT-CNN-GMM and FFT-CNN-GMM-SW subsystems

This subsystem is based on usage of Fast Fourier Transform for obtaining spectrograms. Here we considered two types of feature extraction modules. In the first system we extract quantized grayscale images using the sliding window with 200 window size and 20 length step (we named it FFT-CNN-GMM-SW). In the alternative system we extract one 400 width image from the spectrogram (FFT-CNN-GMM). Similar to previous system high level feature extractor is based on Light CNN implementation.

2.5 LPCC-TV-SVM subsystem

This subsystem is based on usage of Linear Prediction Cepstral Coefficients (LPCC) as informative features. Here we use the standard TV-JFA (Total Variability Joint Factor Analysis) approach for the acoustic space modelling [1]. According to this version of the joint factor anlysis, the i-vector of the Total Variability space is extracted from the whole speaker utterance by means of JFA modification, which is a usual Gaussian factor analyser defined on mean supervectors of the UBM (Universal Background Model) and Total-variability matrix T. In this module the UBM is represented by the 128-component Gaussian mixture model of the described features, and the dimension of the T-matrix was 200. We use the acoustic features obtained for training part of the ASV spoof challenge 2017 database to train both of them. The diagonal covariance UBM was trained by the standard EM-algorithm. These i-vectors are centered and length-normalized.
Normalized i-vectors are then used as input vectors for an SVM classifier. In our system we use the SVM classifier with a linear kernel, which was trained on the i-vectors of the training part of the ASV spoof challenge 2017 database. In SVM training the efficient LIBLINEAR [3] library was used for calculations.

2.6 FFT-CNN-RNN subsystem

This subsystem uses Fast Fourier Transform spectrograms extracted with following parameters: window length is 256, the 192 points are common for each two adjoining windows, the Blackman window function with $\alpha = 0.16$ is used as windowing function.

The end-to-end module in this system consists of Convolutional Neural Network followed by Recurrent Neural Network as described in [6]. Overall architecture is shown in figure 2.

CNN architecture is a reduced version of Light CNN, introduced in [7] with all the max pooling operations being applied using stride 2×1 instead of 2×2. RNN consists of two gated recurrent unit (GRU) blocks: one for the forward pass and one for the backward pass. Last output vectors of both forward and backward passes are concatenated to obtain the common output. Such RNN model is applied to each channel of CNN's output and then results are concatenated. Two fully connected layers with MaxOut activations are used then to obtain 256-dimensional vector, which is reduced by fully connected layer with sigmoidal activation resulting in probability of utterance being spoofed.

2.7 ΔEEMD-CNN-RNN system

This subsystem is completely identical to FFT-CNN-RNN one, but ΔEEMD features are used instead of FFT. Following algorithm is used to obtain ΔEEMD features:

1. Let S_o be the FFT spectrogram of the original signal $x(t)$

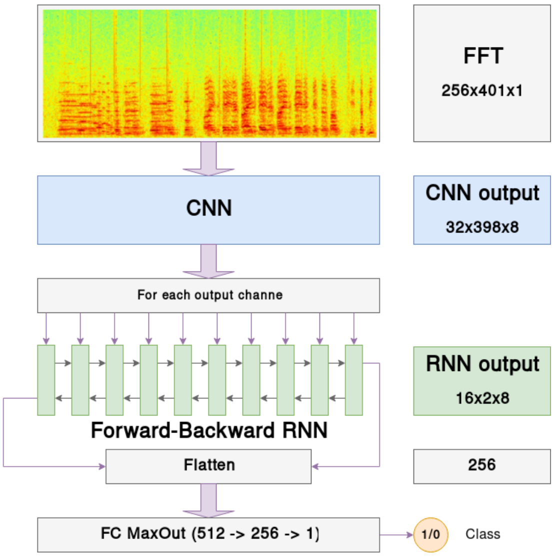

Fig. 2: CNN-RNN architecture

2. Using ensemble empirical mode decomposition (EEMD) with ensemble size of 50 and noise strength equals to $0.1 \cdot \sqrt{\mathrm{Var} x(t)}$ get the first empirical mode $c_1(t)$ of the signal $x(t)$
3. Compute S_r as the FFT spectrogram of the signal $c_1(t)$
4. $S_\Delta = |S_o - S_r|$

libEEMD library [8] is used to build the first empirical mode of the signal.

3 Inhouse databases

For flexible training conditions we collected additional data bases in order to add them to the training dataset. All databases were recorded by modern smartphones. For Inhouse_RSR database we randomly chose 500 utterances as a genuine speech. And for Inhouse base 1 and Inhouse base 2 we used genuine utterances of the dev part from ASVspoof 2017 database. These databases were recorded in different recording environments and with the use of different recording and playback devices. The details of the collected databases are presented in Table 1.

Thus training dataset for the flexible conditions contained:

- ASVspoof 2017 train part
- Inhouse_RSR
- Inhouse base 1

Table 1: Inhouse datasets

Database name	Inhouse_RSR	Inhouse base 1	Inhouse base 2
Recorded database	RSR 2015	genuine' part of 'dev' dataset from ASVspoof 2017	genuine' part of 'dev' dataset from ASVspoof 2017
Playback devices	Sony Xperia Z2 Samsung Galaxy HTC	Sony Xperia Z2 Samsung Galaxy HTC	Sony Xperia Z2 Samsung Galaxy A3 Samsung Galaxy Tablet Xiaomi Redmi 3s
Recording devices	SVEN IHOO MT 5.1	- Dialog Speakers Audio System - SVEN IHOO MT 5.1	- SVEN MA-331 - Panasonic s-xbs bi-wiring system RX-CT990
Recording enviromnent	Office	Office	Home
Number of genuine utterances	500	-	-
Number of spoofing utterances	1496	2807	6072

– Inhouse base 2

And the total replay-to-genuine ratio was around 4.0.

4 Submission systems

For submission we prepared several systems with score-level fusion. Fusion of the subsystems scores was done by the BOSARIS toolkit for MATLAB [4]. Details of the submitted systems are presented in the Table 2.

Our flexible primary system consists of several subsystems trained on the ASVspoof 2017 database and also two systems trained on the extended train database (FFT-CNN-GMM_ext and CQT-CNN-GMM_ext).

References

1. M. Li, X. Zhang, Y. Yan, and S. Narayanan, "Speaker verification using sparse representations on total variability i-vectors," in Interspeech 2011, Aug, 2011
2. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, "Going Deeper with Convolutions", Computer Vision and Pattern Recognition (CVPR) (2015)

Table 2: Submission systems

	Training condition	
Submission	Common	Flexible
Primary	FFT-CNN-GMM-SW + LPCC-TV-SVM + FFT-CNN-RNN	FFT-CNN-GMM + FFT-CNN-GMM-SW + LPCC-TV-SVM + FFT-CNN-RNN + DWT-CNN-GMM-SW + ΔEEMD-CNN-RNN + CQT-CNN-GMM + CQT-CNN-GMM-SW + FFT-CNN-GMM_ext + CQT-CNN-GMM_ext
Contrastive1	FFT-CNN-GMM	FFT-CNN-GMM
Contrastive2	DWT-CNN-GMM-SW + FFT-CNN-GMM + FFT-CNN-GMM-SW + CQT-CNN-GMM + ΔEEMD-CNN-RNN + LPCC-TV-SVM+ FFT-CNN-RNN	ΔEEMD-CNN-RNN + FFT-CNN-RNN

3. LIBLINEAR library `http://www.csie.ntu.edu.tw/~cjlin/liblinear`
4. BOSARIS toolkit `https://sites.google.com/site/bosaristoolkit`
5. M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification antispoofing: Constant Q cepstral coefficients," in Proc. Odyssey, Bilbao, Spain, 2016.
6. C. Zhang; C. Yu; J. H. L. Hansen, "An Investigation of Deep Learning Frameworks for Speaker Verification Anti-spoofing," in IEEE Journal of Selected Topics in Signal Processing , vol.PP, no.99, pp.1-1
7. Wu, X., He, R., Sun, Z., & Tan, T. "A Light CNN for Deep Face Representation with Noisy Labels", arXiv:1511.02683, 2015
8. libEEMD library `https://bitbucket.org/luukko/libeemd`