STC ANTI-SPOOFING SYSTEMS FOR THE ASVSPOOF 2015 CHALLENGE

Sergey Novoselov [1,2], *Alexandr Kozlov*[2], *Galina Lavrentyeva*[1,2],
Konstantin Simonchik [1,2], *Vadim Shchemelinin*[1,2]

[1]ITMO University, St. Petersburg, Russia
[2]Speech Technology Center Ltd., St. Petersburg, Russia
{novoselov, kozlov-a, lavrentyeva, simonchik, shchemelinin}@speechpro.com

ABSTRACT

This paper presents the Speech Technology Center (STC) systems submitted to Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenge 2015. In this work we investigate different acoustic feature spaces to determine reliable and robust countermeasures against spoofing attacks. In addition to the commonly used front-end MFCC features we explored features derived from phase spectrum and features based on applying the multiresolution wavelet transform. Similar to state-of-the-art ASV systems, we used the standard TV approach for probability modelling in spoofing detection systems. Experiments performed on the development and evaluation datasets of the Challenge demonstrate that the use of phase-related and wavelet-based features provides a substantial input into the efficiency of the resulting STC systems. In our research we also focused on the comparison of the linear (SVM) and nonlinear (DBN) classifiers.

Index Terms-spoofing, anti-spoofing, speaker recognition, phase spectrum, wavelet transform, TV, SVM, DBN

1. INTRODUCTION

Information technology plays an increasingly large role in today’s world, and different authentication methods are used for restricting access to informational resources, including voice biometrics. Examples of the use of speaker recognition systems include internet banking systems, customer identification during a call to a call center, as well as passive identification of a possible criminal using a preset “blacklist” [1], [2]. Due to the importance of the information that needs to be protected, requirements for biometric systems are high, including robustness against potential break-ins and other attacks.

Performance of basic technologies in voice biometrics has greatly improved in recent years. For instance, the latest overviews of speaker recognition systems showed that EER is down to 1.5-2% for text-independent [3] and down to 1% for text-dependent [4] speaker recognition systems in various conditions.

With the growth of interest in reliable ASV systems, the development of their spoofing techniques increased tremendously [1]. A multitude of different spoofing methods was proposed in literature. For example, [5] describes methods based on “Replay attack”, “Cut and paste”, “Handkerchief tampering” and “Nasalization tampering”. Speech synthesis approaches [6] are also widely used for spoofing purposes.

Despite the development of new spoofing detection methods, most of ASV spoofing countermeasures presented so far depend on a training dataset related to a specific type of attack, while the nature of the spoofing attack is usually unknown in real practice. Several papers on the ASV system robustness evaluation against spoofing attacks [5], [7], [8] show that it is highly important to develop new anti-spoofing techniques to detect unforeseen spoofing attacks when details of the spoofing attacks are unknown, in order to keep the required EER level. That was the motivation for organizing the ASVspoof Challenge 2015 [9] where spoofing detection methods for known [1] and unknown spoofing types were evaluated. ASVspoof Challenge 2015 was focused on a stand-alone spoofing detection task.

In this paper we describe several spoof detection systems that were proposed for the ASVspoof Challenge 2015.

For participation in the challenge we used the standard Total Variability (TV) approach for statistical modelling of the acoustic features of the speech signal. As a classifier we applied Support Vector Machine (SVM) or, alternatively, Deep Belief Network (DBN).

In the paper we concentrate on researching the most appropriate front-end features for the spoofing detection system we propose. In particular, we investigated acoustic features based on the phase spectrum information and features derived by applying the wavelet transform [10]. The aim of our research was to find the most effective method for detecting unknown spoofing attacks.

The remainder of this paper is organized as follows. Section 2 overviews all the proposed spoof detection systems with a brief description of the subsystems it consists of. Section 3 introduces acoustic feature extraction methods in detail. Experimental work is described in Section 4. Finally, our conclusions are presented in Section 5.

2. OVERALL SYSTEM DESCRIPTION

All our systems consist of three main components (Figure 1):

- Acoustic feature extractor
- TV i-vector extractor
- Classifier

After experiments on the training part of the ASVspoof Challenge 2015 database we decided to include pre-detection as a preliminary step in the spoof detection system. The pre-detector checks whether the input speech signal has zero temporal energy values. In case of zero-sequence the signal is declared to be a spoofing attack, otherwise (or in case of a detection system without the pre-detection step) the speech signal is used as input data for the feature extractor.

Acoustic feature extractors in our systems are combinations of several different acoustic feature extraction methods. All the features we proposed during the ASVspoof Challenge 2015 will be described in details in Section 3.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

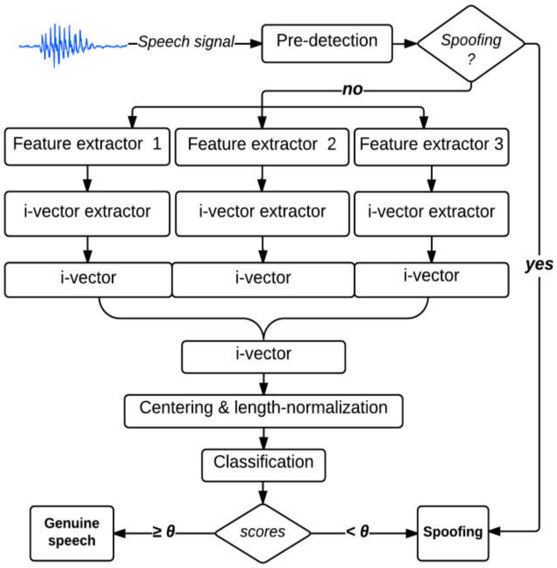


Figure 1: *Scheme of spoof detection system.*

The obtained feature vectors are used by the i-vector extractor to get i-vectors from different feature types. These i-vectors are then concatenated in one common i-vector, which is centered and length-normalized.

Finally, the classifier calculates the resulting score to estimate if the speech signal is a spoofing attack or not (Figure 1).

2.1. Total variability modeling

In our work for the acoustic space modelling we used the standard TV approach, which is the state-of-the-art in speaker verification systems [3], [4], [11]. According to this version of the joint factor analysis, the i-vector of the Total Variability space is extracted by means of Gaussian factor analyser defined on mean supervectors of the Universal Background Model (UBM) and Total-variability matrix T. UBM is represented by the diagonal covariance Gaussian mixture model of the described features.

2.2. SVM-Classification

In our work we used an SVM classifier with a linear kernel. The separating hyperplane was trained in normalized i-vectors space (Figure 1) on the training part of the ASV spoof challenge 2015 database to detect genuine speech phonograms and spoof attacks. In SVM training the efficient LIBLINEAR [12] library was used for calculations in order to achieve the necessary accuracy and computational speed.

2.3. DBN-Classification

Alternatively, we used a classifier based on Deep Belief Network with softmax output units and stochastic binary hidden units [13] (Figure 2). DBN takes normalized i-vectors from the i-vector extractor as input data. We used layer-wise pretreating of the layers by means of Restricted Boltzmann Machines (RBMs) and then applied back-propagation to train the DBN in a supervised way to perform classification.

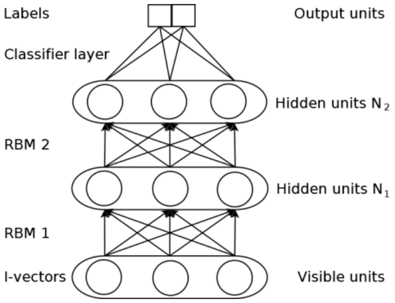


Figure 2: *DBN-Classification.*

3. FRONT-END FEATURES

In this section we describe a number of different acoustic features that were effective for the ASVspoof Challenge 2015 task.

3.1. Amplitude spectral features

As short-term amplitude spectrum acoustic features to be used in the ASV spoofing attack detection system we selected mel frequency coefficients, which were obtained by using the discrete cosine transform for Mel-Frequency Cepstral Coeffitients (MFCC) and principal component analysis for Mel-Frequency Principal Coeffitients (MFPC). These features accurately represent the general characteristics of the vocal tract.

MFCC coefficients represent the short-term power spectrum of the speech signal, based on the application of the discrete cosine transform to the log power spectrum on a nonlinear mel scale of frequency (Figure 3). To derive the MFCC coefficients we used a Hamming window function with the window length equal to 256 and 50% overlap.

We kept the first 12 MFCC coefficients and their first and second-order derivatives as the most informative acoustic features, thereby obtaining a 36-length feature vector.

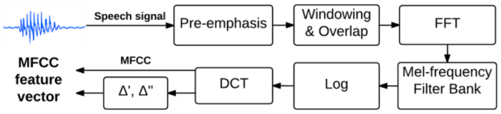


Figure 3: *MFCC feature extractor.*

The MFPC coefficients were obtained similarly to the MFCC coefficients, but using eigenvector basis instead of the discrete cosine basis to achieve decorrelation of the informative acoustic features (Figure 4). Eigenvectors had been calculated on the training part of the ASVspoof Challenge 2015 database.

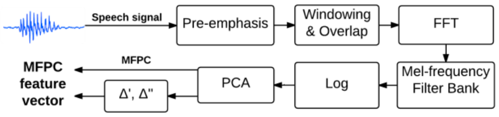


Figure 4: *MFPC feature extractor.*

3.2. Phase-based features

In order to take into account phase information of the speech signal we used the cos-phase features described in [14]. These features were extracted from the phase spectrum, obtained by the Fourier Transform, as follows:

1. The unwrapped phase spectrum was normalized by applying the cosine function to change its range to [-1; 1].
2. Dimensionality reduction was then performed by means of principal component analysis, the basis of which had been calculated beforehand on the training part of the ASVspoof Challenge 2015 database.

Similarly to amplitude spectrum features, we selected only the first 12 coefficients with their first and second-order derivatives to form the resulting feature vector CosPhase Principal Coeffitients (CosPhasePC).

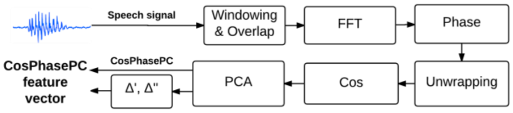
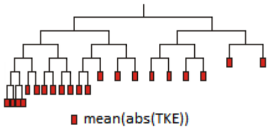


Figure 5: *CosPhasePC feature extractor.*

3.3. Wavelet-based features

We decided to introduce detailed time-frequency analysis of the speech signals in our countermeasures. For this purpose we used front-end features based on applying the wavelet-packet transform [10], adapted to the mel scale (Figure 6). In the wavelet-decomposition we used Daubechies wavelets db4.



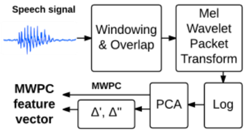


Figure 6: *Wavelet Packet Transform. TKE is a Teager Kaiser energy*

Figure 7: *MWPC feature extractor*

Instead of the classical energy of the frequency sub-bands, here we applied the Teager Keiser Energy Operator (Figure 6). Teager Keiser Energy (TKE) is more informative than classical sample energy. Moreover, it is a noise-robust parameter for speech signal [15]. We used the following equation for TKE evaluation:

$$\Psi\big(s(t)\big) = s(t)^2 - s(t-1)s(t+1) \qquad (1)$$

where $s(t)$ is the output signal temporal sample of the considered sub-band.

For extracted features decorrelation, similarly to 3.1 and 3.2, we consistently applied projection on the eigenvector basis to derive 12 coefficients. We called these features Mel Wavelet Packet Coefficients (MWPC) for short. Equivalently to 3.1 and 3.2, here we observe MWPC with its first and second-order derivatives (Figure 7).

To derive the MWPC coefficients we also used a Hamming window function with the window length equal to 256 and 50% overlap.

4. EXPERIMENTS

For training all parameters of our anti-spoofing systems we used training dataset. All the experiments, described below, were performed on the development dataset. According to the challenge conditions [9], the training and development datasets contained 5 spoofing attacks of types S1-S5: S1,S2,S5 were voice conversion algorithms and S3, S4 were HMM-based speaker-adapted speech synthesis methods.

For example, Figure 8 presents LDA projections of the MWPC based i-vectors for the subset of the development dataset on the 3 principal components P_1, P_2, P_3, evaluated on the training dataset.

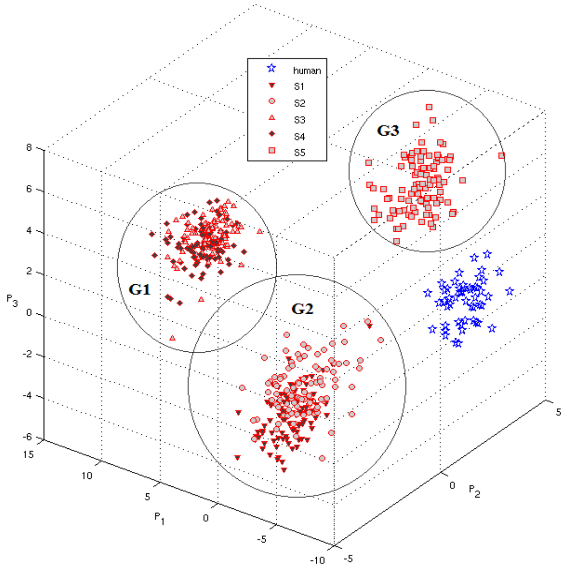


Figure 8: *The LDA projections of MWPC based i-vectors on principal components P_1, P_2, P_3.*

In Figure 8 we can see that that human class is well separated from the spoofing classes. This allows us to argue that integration of SVM linear classification methods can be efficient for solving the spoofing detection task in this space. Note that in this feature space 3 groups of spoofing classes can also be easily discriminated: G1 is the group of HMM-based speech synthesis (S3+S4); G2 is the group of simple Voice conversion techniques (S1+S2); G3 represents Festvox voice conversion method.

Table 1 demonstrates resulting EER estimates (%) of the TV-SVM based system with different front-end features, described in Section 3, obtained on the development dataset. These results were obtained with the use of 256-component UBM and 200 dimensional i-vector.

Table 1. *Experimental results for the TV-SVM systems for different types of spoofing algorithms, EER (%).*

Features	***Spoofing algorithm***					
	S1	***S2***	***S3***	***S4***	***S5***	***All***
MFCC	0.38	2.13	0.36	0.39	1.48	1.14
MFPC	0.13	0.29	0.09	0.09	0.37	0.23
CosPhasePC	0.13	0.20	0.04	0.05	0.23	0.15
MWPC	**0.03**	**0.11**	**0.00**	**0.00**	**0.08**	**0.05**

The results show that standard MFCC features are inferior in comparison with other front-end features by EER estimates. It should be mentioned that substantial EER improvement was achieved on MFPC for all spoofing techniques by PCA basis applying in MFCC instead of discrete-cosine transform. CosPhasePC implementation slightly reduces EER compared with MFPC. Features based on the multiresolution wavelet transform outperform all other proposed features, reaching 0.05% EER for all known attacks. Note that during our experiments we determined that using the TKE operator in MWPC demonstrates better results than classical energy.

4.1. Fusion

In this work we explored fusion of systems based on different front-end features in the i-vector space. All i-vectors derived from different extractors were concatenated in one common i-vector, as shown in Figure 1. Results for different fusion systems confirm that for all combinations of features EER estimations for a known spoofing attack is close to 0% (Table 2). However, the zero error of spoofing detection for the development dataset was achieved only by implementing the combination of CosPhasePC and MWPC features.

Table 2. *Experimental results for the fused TV-SVM systems for all spoofing algorithms, EER (%)*

Features	*All known attacks*
MFCC, MFPC, CosPhase	0.03
MFCC,MFPC, MWPC	0.03
MFCC,MWPC,CosPhase	0.00
MFPC,MWPC, CosPhase	0.00

4.2. Classifiers

Experiments of the fused TV systems with the combination of MFCC, MFPC and CosPhasePC feature extractors based on SVM and DBN classifiers showed that SVM-based system achieved 0.03% EER on the development dataset, which is better than 0.04% EER of DBN-based system.

5. EVALUATION RESULTS

Based on the results of our experiments on the development dataset, we decided to propose 3 systems according to the common condition of the ASVspoof Challenge 2015.

Our **Primary** system was implemented according to the scheme on Figure 1 with a pre-detection step. MFCC, MFPC and CosPhasePC were selected to be the informative acoustic features in this system. The UBM was represented by a 1024-component Gaussian mixture model of the described features, and the dimension of the TV space was 400. Here we used SVM for classification.

In our **Contrastive 1** system, in contrast to the Primary system, a pre-detector was not used, and MWPC features were used instead of MFCC.

Our third system **Contrastive 2** also did not use a pre-detector, and instead of the SVM classifier it applied a nonlinear DBN classifier. In this system, in order to avoid overfitting, we reduced the UBM component number to 256 for all feature types, while the dimension of the TV space was reduced to 200.

In addition to known spoofing attacks, the evaluation dataset contained spoofing attacks of 5 unknown types S6-S10. Table presents the evaluation results.

Table 3. *Evaluation results of the STC systems, EER (%).*

System	*Known attacks*	*Unknown attacks*	*All*
Primary	**0.008**	**3.922**	**1.965**
Contrastive 1	0.009	4.891	2.450
Contrastive 2	0.017	6.162	3.090
Primary (no pre-detection)	0.008	5.151	2.579

In spite of the good results of the STC systems for known attacks, results obtained for unknown attacks are much worse: our best primary system reached 3.92% EER. This observation suggests the necessity to improve countermeasures to achieve robust performance on unknown attacks.

The primary system showed the best results, in particular due to the energy pre-detection step. To confirm this suggestion we compared this result with EER for primary system without pre-detector (Table 3). This pre-detector will not be useful in case of channel effects or additive noise. That is a significant limitation of this anti-spoofing system.

Unlike the Primary system, the Contrastive 1 system did not use the pre-detector, and it demonstrated relatively good performance according to the challenge results. We see the reason for that in the MWPC features. The advantage of these features is the wavelet-transform, which makes it possible to produce detailed multiresolution signal analysis. That provides an additional decrease of EER in the spoofing detection task.

Results for the Contrastive 2 system (4), which used the DBN classifier, turned out to be the worst. In this system we probably failed to avoid the effects of the stronger overfitting on the training dataset, in comparison with SVM. According to these results it can be suggested that it is better to use linear SVM classifier in the proposed system (Figure 1).

Table 4. *Evaluation results for STC systems for different types of spoofing algorithms, EER (%).*

Spoofing algorithm	*System*		
	Primary	*Contrastive 1*	*Contrastive 2*
S1	0.004	0.005	0.000
S2	0.022	0.022	0.058
S3	0.000	0.000	0.000
S4	0.000	0.000	0.000
S5	0.013	0.020	0.029
S6	0.019	0.024	0.046
S7	0.000	0.007	0.000
S8	0.015	0.014	0.124
S9	0.004	0.006	0.005
S10	19.571	24.401	30.636

6. CONCLUSIONS

In this paper we produce comprehensive investigation of the feature spaces applicability and effectiveness of different classifiers in solution of the spoofing detection task for ASVspoof Challenge 2015. The submitted STC systems were based on TV modelling in the space of different features and a linear SVM-based classifier or a nonlinear DBN-based classifier.

During our research we tested different front-end features, including features derived from phase spectrum and features obtained by applying wavelet-transform, in order to determine reliable and robust countermeasures in the anti-spoofing task. Experiments performed on the development and test datasets of the Challenge demonstrate that using phase-based and wavelet-based features provides a substantial input into the efficiency of the resulting STC systems. Our best result on the test dataset is 1.965% EER for all spoofing attacks.

7. ACKNOWLEDGMENTS

This work was partially financially supported by the Government of Russian Federation, Grant 074-U01.

8. REFERENCES

[1] 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, no. 0, pp. 130– 153, 2015.

[2] Matveev Yu. N., “Biometric technologies of person identification by voice and other modalities”, Vestnik MGTU. Priborostroenie, Special Issue 3(3) “Biometric Technologies”, pp. 46-61, 2012 (rus).

[3] A.Kozlov, O. Kudashev, Yu. Matveev, T. Pekhovsky, K. Simonchik and A. Shulipa, “Speaker recognition system for the NIST SRE 2012”, SPIIRAS Proceedings, 2, 2013.

[4] S. Novoselov, T. Pekhovsky, K. Simonchik “STC Speaker Recognition System for the NIST i-Vector Challenge” in Proc. Odyssey 2014 - The Speaker and Language Recognition Workshop, 2014.

[5] Villalba E., Lleida E., “Speaker verification performance degradation against spoofing and tampering attacks”, in Proc. of the FALA 2010 Workshop, pp. 131–134, 2010.

[6] Pavel Chistikov, Dmitriy Zakharov and Andrey Talanov, “Improving Speech Synthesis Quality For Voices Created From an Audiobook Database. A. Ronzhin et al. (Eds.)”, SPECOM 2014, LNAI 8773, pp. 276–283, 2014.

[7] Shchemelinin V., Topchina M., Simonchik K., “Vulnerability of Voice Verification Systems to Spoofing Attacks with TTS Voices Based on Automatically Labeled Telephone Speech”, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 8773, No. LNAI, pp. 475–481, 2014.

[8] Sébastien Marcel, Mark S. Nixon,Stan Z. Li, “Handbook of Biometric Anti-spoofing: Trusted Biometrics Under Spoofing Attacks”, Springer, 2014.

[9] Zhizheng Wu, Tomi Kinnunen, Nicolas Evans, Junichi Yamagishi, Cemal Hanilci, Md Sahidullah, Aleksandr Sizov, “ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge”, Interspeech, 2015.

[10] Stéphane Mallat, A Wavelet Tour of Signal Processing, 3rd ed., Academic Press, Dec. 2008.

[11] N. Dehak et al., “Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification”, in Proc. Interspeech, 2009, pp. 1559-1562.

[12] LIBLINEAR: A Library for Large Linear Classification. http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[13] Hinton, G. E., Osindero, S., & Teh, Y, “A fast learning algorithm for deep belief nets”, Neural Computation, Vol. 18, pp. 1527–1554, Jul. 2006.

[14] Zhizheng Wu, Eng Siong Chng, Haizhou Li, “Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition”, in Proc. Interspeech, 2012.

[15] Ying G. S., Mitchell C.D., and Jamison L.H. “Endpoint detection of isolated utterances based on modified teager energy measure”, ICASSP, 1993, pp. 732-735.