A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 187

On the Generalization of Fused Systems in Voice Presentation Attack Detection

André R. Gonçalves[1], Pavel Korshunov[2], Ricardo P.V. Violato[1],
Flávio O. Simões[1], Sébastien Marcel[2]

Abstract: This paper describes presentation attack detection systems developed for the Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017). The submitted systems, using calibration and score fusion techniques, combine different sub-systems (up to 18), which are based on eight state of the art features and rely on Gaussian mixture models and feed-forward neural network classifiers. The systems achieved the top five performances in the competition. We present the proposed systems and analyze the calibration and fusion strategies employed. To assess the systems' generalization capacity, we evaluated it on an unrelated larger database recorded in Portuguese language, which is different from the English language used in the competition. These extended evaluation results show that the fusion-based system, although successful in the scope of the evaluation, lacks the ability to accurately discriminate genuine data from attacks in unknown conditions, which raises the question on how to assess the generalization ability of attack detection systems in practical application scenarios.

Keywords: Presentation attack detection, spoofed speech, cross-database evaluation.

1 Introduction

Presentation (or replay) attacks can be considered as one of the major obstacles preventing the adoption of speaker recognition in practical applications. This type of attack is relatively easy to perform. If an attacker has access to a speech sample from a target user, he/she can replay it using a loudspeaker or a smartphone to the biometric system during the authentication process. The ease of perpetration and the fact that no technical knowledge of the biometric system is required makes the presentation attack one of the most common practical attacks. Despite the severity of the problem, researchers started to develop effective presentation attack detection mechanisms only in the last few years [SKH15].

One of the main challenges in Presentation Attack Detection (PAD) is to find a set of features that allows systems to effectively distinguish speech signals that were directly emitted by a human vocal apparatus from those reproduced by a replay device such as a loudspeaker or a smartphone. Several audio descriptors originally proposed for speaker verification and speech recognition have also been studied in the context of PAD systems [SKH15] (and references there in). Features specifically designed for anti-spoofing systems were the focus of recent research [CRS07, TDE16, MMDM16].

[1] CPqD, Brazil. {andrerg,rviolato,simoes}@cpqd.com.br
[2] Idiap Research Institute, Switzerland. {pavel.korshunov,sebastien.marcel}@idiap.ch

Generalization ability of PAD systems has been assessed recently with [TDE17] showing the degradation in performance when specific features optimized using one database are used unchanged on another database. In [PSS17], cross-database experiments demonstrated the inability of current techniques to deal with unforeseen conditions. However, it did not include strict presentation attacks, which can be considered one of the hardest attack to be detected. The authors of [KM16, KM17] focused on presentation attacks in cross-database and cross-attack scenarios, and concluded that current state of the art PAD systems do not generalize well, with especially poor performance on presentation attacks.

In this paper, we present two PAD systems developed for the Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) [Ki17b]. The submitted systems are essentially ensembles of several sub-systems composed of state-of-the-art features in PAD systems and two well known classifiers: Gaussian Mixture Models (GMM) and feed-forward neural networks. Calibration and fusion strategies were used to effectively integrate these sub-systems into a possibly more robust PAD systems. We discuss and compare three different fusion strategies and investigate their performances on the ASVspoof 2017 database, as well as, by using an unrelated and larger database recorded in Portuguese language: BioCPqD-PA [Vi13] database.

2 Database and Protocol

Two different databases were used: ASVspoof 2017 and BioCPqD-PA, containing genuine and spoofed recordings. The protocol defined in the ASVspoof challenge splits the database into three subsets, while BioCPqD-PA is used as just one set. Table 1 summarizes both datasets. The databases and protocols are described in the following subsections.

Tab. 1: Number of speakers and utterances in ASVspoof 2017 and BioCPqD-PA databases.

	ASVspoof 2017			BioCPqD-PA
	train	*dev*	*eval*	-
# speakers	10	8	NA	222
# genuine	1,508	760	1,298	27,253
# spoofed	1,508	950	12,008	42,768
# total	3,016	1,710	13,306	70,021

2.1 ASVspoof 2017

The ASVspoof 2017 contest focuses on presentation attacks. To this end, the challenge makes use of the RedDots corpus [Le15] and a replayed version of the same data [Ki17a]. While the former serves as genuine samples, the latter is used as spoof samples, collected by replaying a subset of the original RedDots corpus utterances using different loudspeakers and recording devices, in different environments, through a crowdsourcing approach.

The database was split into three subsets: *train* for *training*, *dev* for *development*, and *eval* for *evaluation*. It was also allowed to use both *train* and *dev* subsets to train the final system

for score submission. The evaluation metric adopted was the Equal Error Rate (EER) and there was no need for participants to provide a decision threshold. The only restriction concerning the score was that higher scores should favor the genuine hypothesis and lower scores the spoof hypothesis. A detailed description of the contest can be found in the challenge evaluation plan [Ki17b].

2.2 BioCPqD-PA

BioCPqD-PA [Vi13] is a proprietary database that contains videos (image and audio) of participants recorded on different devices (laptops and smartphones) and environments. All recordings are in Portuguese language. The recordings (genuine audios) are from 222 speakers, collected with 4 different laptops, in 3 distinct environments, and during 5 recording sessions. In each session, 27 utterances with variable content were recorded. The total of genuine audios is 27,253.

To create the spoof attacks, a subset of these recordings (1,782 utterances sampled from all speakers' utterances in such a way that all speakers were represented in the subset) were replayed in an acoustically isolated room, using 3 different microphones and 8 different loudspeakers, resulting in 24 configurations. Then, the total number of spoofed recordings is 42,768 samples (see Table 1). In the cross-database experiments, BioCPqD-PA was used as one set. Therefore, systems tuned and trained on the ASVspoof 2017 database (following its protocol) were evaluated on the entire BioCPqD-PA, and, likewise, a system with the same configuration was trained on BioCPqD-PA and tested on ASVspoof 2017.

3 Description of the submitted PAD systems

In this section, we describe the components that constitute the two submitted PAD systems referred to as *System-1* and *System-2* in the rest of the paper.

3.1 Features

We evaluated the performance of the following features previously investigated in the context of spoofing attacks with synthetic speech: MFCC, IMFCC, RFCC, LFCC, PLP-Cepstral, SCMC, and SSFC. The use of these features was inspired by [SKH15]. Feature implementations are available online[3], which contribute to reproducibility of results. Other features were considered, such as CQCC [TDE16], PNCC [KS16], and GFCC [VA12], but in our previous tests they did not improve the performance of the jointly fused PAD system.

Features are extracted from 20ms speech frames with 50% overlap. All features are based on short-term power spectrum and were considered 20 coefficients along with their delta and delta-delta dynamic coefficients.

[3] `http://cs.joensuu.fi/~sahid/codes/AntiSpoofing_Features.zip` and in Bob framework `https://www.idiap.ch/software/bob/`

3.2 Classifiers

Two distinct classifiers were employed: a traditional 2-class Gaussian Mixture Model (GMM) classifier, where two 512 components GMM were trained (10 EM iterations), one for each class, and a Feed Forward Neural Network (FFNN), with the following architecture: Input $d \times 1 \rightarrow$ fully connected (12 neurons ReLU) $\rightarrow$ Batch Normalization $\rightarrow$ Dropout ($p = 0.5$) $\rightarrow$ fully connected (64 neurons ReLU) $\rightarrow$ Dropout ($p = 0.5$) $\rightarrow$ Sigmoid output. The cross-entropy cost function was minimized via Stochastic Gradient Descent with learning rate equals to 1e-4 with Nesterov's acceleration.

For 2-class GMM implementation, we used the system provided by the organizers with the baseline system[4] and the implementation in Bob framework[5], while FFNN was implemented in python using theano/keras[6] framework.

It is important to point out that testing different classification techniques was beyond the scope of our work for this evaluation. Therefore, a lot of space remains for assessing the use of more elaborate classifiers for PAD.

3.3 Calibration and Fusion

We focus on a score level fusion due to its relative simplicity and evidence that it leads to a better performance. The score-fusion is performed by combining scores from each of the N systems into a new feature vector of length N that needs to be classified. For classification we consider three different algorithms: (*i*) a logistic regression (LR), (*ii*) a multilayer perceptron (MLP), and (*iii*) a simple average function (Avg), which is taken on scores of the fused systems. For LR and MLP fusion, the classifier is pre-trained on the score-feature vectors from the training set.

When analyzing, comparing, and especially fusing PAD systems, it is important to have calibrated scores. Raw scores can be mapped to log-likelihood ratio scores with logistic regression, and an associated cost of calibration C_{llr} together with a discrimination loss C_{llr}^{min} are then used as application-independent performance measures of calibrated PAD or ASV systems. Calibration cost C_{llr} can be interpreted as a scalar measure that summarizes the quality of the calibrated scores. A well-calibrated system has $0 \leq C_{llr} < 1$ and produces well-calibrated likelihood ratio. Discrimination loss C_{llr}^{min} can be viewed as the theoretically best C_{llr} value of an optimally calibrated systems. We refer to [Ma14] for a discussion on the score calibration and C_{llr} and C_{llr}^{min} metrics.

3.4 Submitted systems

The two submitted PAD systems are essentially ensembles of different combinations of features and classifiers. Table 2 shows the set of sub-systems and the fusion method used

[4] `http://www.ASVspoof.org/data2017/baseline_CM.zip`
[5] `https://gitlab.idiap.ch/bob/bob.bio.gmm`
[6] Theano: `https://github.com/Theano/Theano` and Keras: `https://keras.io/`

for each PAD system. Features are presented with a subscript '*all*' or 'Δs', where '*all*' means that all static and dynamic (delta and delta-delta) features were used, while 'Δs' indicates that only the dynamic features were considered. The choice of the set of sub-systems was based on their performances measured on contest's *dev* set prior to the sub-mission.

Tab. 2: Description of the submitted systems: *System-1* and *System-2*.

	System-1	System-2
Sub-systems	**GMM** with: RFCC$_{all}$, RFCC$_{\Delta s}$, LFCC$_{all}$, LFCC$_{\Delta s}$, MFCC$_{all}$, MFCC$_{\Delta s}$, IMFCC$_{all}$, MFCC$_{\Delta s}$, SSFC$_{all}$, SSFC$_{\Delta s}$, SCMC$_{all}$, SCMC$_{\Delta s}$ **FFNN** with: IMFCC$_{all}$, LFCC$_{all}$, MFCC$_{all}$, PLP-Cepstral$_{all}$, RFCC$_{all}$, SCMC$_{all}$	**GMM** with: RFCC$_{all}$, RFCC$_{\Delta s}$, LFCC$_{all}$, LFCC$_{\Delta s}$, MFCC$_{all}$, MFCC$_{\Delta s}$, IMFCC$_{all}$, IMFCC$_{\Delta s}$, SSFC$_{all}$, SSFC$_{\Delta s}$, SCMC$_{all}$, SCMC$_{\Delta s}$
Fusion	Logistic Regression	Logistic Regression

4 Results on the ASVspoof2017 database

Table 3 shows the performance of the submitted systems in terms of EER, both for the *dev* and the *eval* sets. The results obtained for the *dev* set are based on the systems trained exclusively on the *train* set of ASVspoof2017 database, while to obtain the results for *eval* set, the systems were trained on the aggregated set: *train+dev*.

Additionally, the table shows the results of baseline system provided by the challenge organizers, which is based on CQCC front-end and 2-class GMMs back-end. *Best individual* system corresponds to a single IMFCC-based sub-system trained using GMM, which demonstrated the best performance during pre-submission evaluations. A detailed analysis of the results can be found in [Ki17b], where the results from all participants are compared.

Tab. 3: EER results for the systems submitted to ASVspoof2017, the baseline system, and the best individual model (GMM with IMFCC). The performance degradation in the Eval set is possibly due to the presence of unknown attacks. Ensemble models (System-1 and System-2) are more robust than individual models on the unseen conditions in the Eval set. Best results are highlighted.

	System-1	System-2	Best individual	Baseline
Dev (train only)	**4.09**	4.32	4.86	11.17
Eval (train+dev)	**14.31**	14.93	29.41	24.65

The only difference between baseline and best individual system is the features used, as the classifier is the same. An interesting result is the one obtained with best individual system. While on the *dev* set it provides comparable performance to the fusion-based systems, on the *eval* set it performs dramatically worse.

5 Cross-database analysis

To asses the real ability of the systems trained on the challenge database we applied them to the completely unrelated BioCPqD-PA database.

Tab. 4: EER results for the cross-database experiments: system trained on ASVspoof 2017 (*train+dev*) and tested on BioCPqD-PA, and system trained on BioCPqD-PA and tested on ASVspoof 2017 (*eval*). Best results are highlighted.

	System-1			System-2			Best
	Avg	LR	MLP	Avg	LR	MLP	individual
ASVspoof (train+dev) → BioCPqD-PA	23.35	**21.35**	22.34	22.23	**21.28**	22.41	37.24
BioCPqD-PA → ASVspoof (eval)	31.86	**26.58**	30.77	27.74	27.96	28.37	27.77

Table 4 shows that the systems trained on the ASVspoof2017 challenge database (*train+dev*) and tested on BioCPqD-PA database led to twice larger EER compared to when the same systems are evaluated on the *eval* set of ASVspoof2017 (see Table 3). This finding confirms the limited generalization power of the systems. The performance degradation in cross-database experiments is not unprecedented: it has been observed in previous anti-spoofing evaluations [TDE17, PSS17, KM16].

Three different fusion methods using Average, LR, and MLP algorithms were tested with comparable performances. LR led to a slightly better performance, especially for *System-1* trained on BioCPqD-PA database and evaluated on ASVspoof. Comparing the best individual sub-systems against fused systems, although fusion did not improve results for systems trained on BioCPqD-PA database, there is a significant improvement when it is trained on ASVspoof database. Thus, we can reason that, in practice, when the scenario is unknown, fusion add robustness to the system performance.

Observing the non-negligible difference between the two crossing possibilities in Table 4, one can arguably say that training data diversity matters. While ASVspoof database has few speakers (only male) and a limited number of utterances, it contains presumably more diverse conditions (devices and recording environments) than BioCPqD-PA, due to the crowdsourcing data collection. On the other hand, BioCPqD-PA is larger, both in terms of speakers and number of utterances, but recording conditions are more restricted.

6 Discussion

In every challenge, such as ASVspoof or NIST SRE (Speaker Recognition Evaluation[7]), the discussion about the provided speech databases emerges. Todisco et al. [TDE17] discuss the problem of selecting the features set based on results in one database and using it on another set, pointing out the resulting performance degradation. Based on our experiments, we raise another question regarding the generalization capability of systems to completely unseen conditions (including different language). Such situation is more likely to happen in practical PAD systems, where the system is trained on a given database and the attacks come from completely unknown conditions.

[7] https://www.nist.gov/itl/iad/mig/speaker-recognition

One should note that our cross-database experiments were designed for an extremely mismatched situation, when even the language is different between databases. It is expected that a PAD system should not be sensitive to language mismatch, however that might not be the case in practice, as most speech features represent acoustic properties of speech that are indeed affected by the language spoken. This has been a concern for the speaker recognition community as well: the effect of language mismatch has been evaluated in speaker recognition tasks within NIST SRE along the years.

Training a system with good generalization capability might require a larger and more diverse database. Modern algorithms based on deep learning [GBC16] approaches, for instance, which have proven to beat standard approaches in different kinds of tasks, such as speech recognition and computer vision, need massive amounts of data to provide state-of-the-art performance. In cases when the acquisition of such an amount of data is unfeasible, data augmentation strategies, such as [GUNY15], should be considered.

Another point that leads to a controversy is the use of so-called *megafusion* strategies. Although the fusion of many systems, sometimes more than a dozen (e.g., the submitted *System-1* is a fusion of 18 systems), usually leads to a better performance, its practical use is questionable. Megafusion has also been frequently used for the speaker recognition task, holding the current state-of-the-art results. However, its computational burden makes it unacceptable in practical cases, specially when system's response time is crucial.

7 Conclusions

We presented the attack detection systems developed for the Automatic Speaker Verification Spoofing and Countermeasures Challenge. The two systems achieved top five error rate (in terms of equal error rate) among 48 participants. In addition, experiments are expanded to cross-database scenario (supposedly closer to a realistic application), using BioCPqD-PA, a different unrelated database. In these experiments, a significant degradation in performance of the submitted attack detection systems is observed, highlighting the lack of generalization ability of such systems.

To improve performance, other classifiers, such as support vector machine, random forest, and deep neural networks (DNNs), need to be tested in the future. As high-generalization capability classifiers such as DNNs require a large amount of supervised training data, new data collections or data augmentation strategies will also be considered in future works. Other features specifically designed for presentation attack also need to be investigated.

Acknowledgements

This work was partially funded by Norwegian SWAN project, EU H2020 project TeSLA, and Swiss Center for Biometrics Research and Testing.

References

[CRS07] Chakroborty, J.S.; Roy, A.; Saha, G.: Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks. International Journal of Signal Processing, 4(2):114–122, 2007.

[GBC16] Goodfellow, I.; Bengio, Y.; Courville, A.: Deep Learning. MIT Press, 2016. `http://www.deeplearningbook.org`.

[GUNY15] Gonçalves, A. R.; Uliani Neto, M.; Yehia, H. C.: Accelerating replay attack detector synthesis with loudspeaker characterization. In: 7th Symposium of Instrumentation and Medical Images / 6th Symposium of Signal Processing of UNICAMP. 2015.

[Ki17a] Kinnunen, T.; Sahidullah, M.; Falcone, M.; Costantini, L.; Hautamäki, R. G.; Thomsen, D.; Sarkar, A.; Tan, Z.H.; Delgado, H.; Todisco, M.; Evans, N.; Hautamäki, V.; Lee, K.A.: RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research. In: ICASSP. pp. 5395–5399, 2017.

[Ki17b] Kinnunen, T.; Sahidullah1, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.A.: The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In: Interspeech. 2017.

[KM16] Korshunov, P.; Marcel, S.: Cross-database evaluation of audio-based spoofing detection systems. In: Interspeech. pp. 1705–1709, 2016.

[KM17] Korshunov, P.; Marcel, S.: Impact of Score Fusion on Voice Biometrics and Presentation Attack Detection in Cross-Database Evaluations. IEEE Journal of Selected Topics in Signal Processing, 11(4):695–705, June 2017.

[KS16] Kim, C.; Stern, R. M.: Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(7):1315–1329, 2016.

[Le15] Lee, K.; Larcher, A.; Wang, G.; Kenny, P.; Brümmer, N.; van Leeuwen, D. A.; Aronowitz, H.; Kockmann, M.; Vaquero, C.; Ma, B.; Li, H.; Stafylakis, T.; Alam, M. J.; Swart, A.; Perez, J.: The reddots data collection for speaker recognition. In: Interspeech. pp. 2996 – 2091, 2015.

[Ma14] Mandasari, M. I.; Günther, M.; Wallace, R.; Saeidi, R.; Marcel, S.; van Leeuwen, D. A.: Score calibration in face recognition. IET Biometrics, 3(4):246–256, 2014.

[MMDM16] Muckenhirn, H.; Magimai-Doss, M.; Marcel, S.: Presentation Attack Detection Using Long-Term Spectral Statistics for Trustworthy Speaker Verification. In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). Darmstadt, Germany, pp. 1–6, Sept 2016.

[PSS17] Paul, D.; Sahidullah, M.; Saha, G.: Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora. In: ICASSP. pp. 2047–2051, 2017.

[SKH15] Sahidullah, M.; Kinnunen, T.; Hanilçi, C.: A Comparison of Features for Synthetic Speech Detection. In: Interspeech. pp. 2987 – 3000, 2015.

[TDE16] Todisco, M.; Delgado, H.; Evans, N.: A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients. In: Odyssey, The Speaker and Language Recognition Workshop. pp. 283–290, 2016.

[TDE17] Todisco, M.; Delgado, H.; Evans, N.: Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Computer Speech & Language, 2017.

[VA12] Valero, X.; Alias, F.: Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. IEEE Transactions on Multimedia, 14(6):1684–1689, 2012.

[Vi13] Violato, R.P.V.; Neto, M. Uliani; Simões, F.O.; Pereira, T.F.; Angeloni, M.A.: BioCPqD: uma base de dados biométricos com amostras de face e voz de indivíduos brasileiros. Cadernos CPqD Tecnologia, 9(2):7–18, 2013.