Influence of the attack conditions on countermeasures for Automatic Speaker Verification

Maxime Baelde, Nathan Souviraà-Labastie, Raphaël Greff

Influence of the attack conditions on countermeasures for Automatic Speaker Verification

Maxime Baelde[1], *Nathan Souviraà-Labastie*[1], *Raphaël Greff*[1]

[1]A-Volute SAS, 59491 Villeneuve d'Ascq, France

name.surname@nahimic.com

Abstract

The ASVSpoof challenges goal is to evaluate countermeasures to spoof attacks on automatic speaker verification systems. We first analyze in more details the results of the baseline systems provided by the organization and unveil several weaknesses for some types of attack. In particular for the *physical access* (PA) task, replay attacks with low reverberation time and/or high quality of the replay device are problematic. Based on this observation, we propose several improvements. Firstly, a specific learning targeting the problematic types of attack. Secondly, a new type of feature enhancing the reverberation. Thirdly, a Deep Neural Network with more modelling capability. On the development set of the PA task, each proposed improvements show results amelioration for the targeted types of attack. Furthermore, the ensemble systems based on this proposed improvements show great overall results amelioration compared to the baseline (0.140 vs 0.193 min t-DCF). However, the amelioration is less encouraging on the evaluation set (0.225 vs 0.245 min t-DCF), thus raising the question of over-fitting as the development set and the train set are similar.

Index Terms: speech recognition, antispoofing, late fusion, *feature engineering*

1. The 2019 ASVSpoof challenge

The ASVSpoof challenge[1] consists in building anti-spoofing systems, also called countermeasures (CM), to be used in tandem with Automatic Speaker Verification (ASV) systems. The 2015 ASVSpoof challenge [1] introduced the *Logical Access* (LA) task which consists in recognizing whether a speaker utterance is produced by a real human or by a voice synthesis system. The 2017 ASVSpoof challenge [2] introduced the *Physical Access* (PA) task which consists in recognizing whether a speaker utterance is coming from a real human or from a playback device (such as a smartphone). The 2019 edition of the challenge is the continuation of the two previous ASVSpoof challenge. The organizers [3] provide a training set of examples used for model training, and a development set used for models self-evaluation and hyperparameters tuning. An additional evaluation set (for which the participants do not have access to the ground truth) is provided to rank the different submissions to the challenge and to drive the participants to build systems capable of generalization.

1.1. The LA task, speech synthesis detection

The LA task comprises six speech spoofing systems used to generate synthesized voice. The SS_1, SS_2 and SS_4 speech spoofing systems use neural network acoustic models and respectively WaveNet, conventional vocoder and convention vocoder using Merlin[2]. The US_1 speech spoofing system is a unit-selection system using MaryTTS[3]. Finally, the VC_1 and VC_4 speech spoofing systems are respectively neural network based and transform function based voice conversion system.

1.2. The PA task, replay attack

The PA task comprises five categories of conditions[4] related to the replay attack, each category having three possible ranges available in Table 1. First, the *environment* which consists in three categories: the room size ($\mathcal{S}$), the reverberation time T60 (R), the talker-to-asv distance ($\mathcal{D}_s$). Second, the *attack* which consists in two categories: the attacker-to-talker distance ($\mathcal{Z}$) and the replay device quality (Q). Details can be found in [3].

Table 1: *PA task. Ranges of the environment (first 3 lines) and attack (last 2 lines) parameters [3].*

	Ranges →	a/A	b/B	c/C
Categories	$\mathcal{S}$ (m^2)	2-5	5-10	10-20
	R (ms)	50-200	200-600	600-1000
	$\mathcal{D}_s$ (cm)	10-50	50-100	100-150
	$\mathcal{Z}$ (cm)	10-50	50-100	>100
	Q	Perfect	High	Low

1.3. The EER and t-DCF evaluation metrics

The *equal error-rate* (EER) [4] point is a compromise between benefits (True Positive) and losses (False Positive). It represent the True Positive Rate (TPR) against False Positive Rate (FPR) for different threshold. However EER metric has some disadvantages. Firstly, EER is ill-suited for unbalanced problem. Secondly, EER is not a reliable predictor of performance when several systems are used in tandem. The ASV system can only discriminate between target (1) and spoofing impostor (3) trials and nontarget trials (2) and CM systems are designed to distinguish genuine speech ((1) and (2)) from spoofed speech (3). In so, different combinations of ASV and CM systems can change the EER.

The *tandem Detection Cost Function* (t-DCF) [4] calculates the cost of detection error for the tandem system ASV-CM. The t-DCF is based on four situations: ASV system rejecting a target trial, ASV system accepting a nontarget trial, CM rejecting a human trial and CM accepting a spoof trial. For each situation, we multiply the cost of detection error, prior probability and error probability and the four products are summed to define the t-DCF. Since the t-DCF takes into account both the ASV and CM systems, it is used as the ranking metric for the ASVSpoof 2019 challenge.

[1]http://www.asvspoof.org/

[2]https://github.com/CSTR-Edinburgh/merlin

[3]http://mary.dfki.de

[4]The notations is derived from the challenge description [3].

2. Weakness of the baseline

The challenge organizers defined two baseline systems (LFCC-GMM and CQCC-GMM)[5] that previously demonstrated the best performances on the previous 2015 and 2017 challenges corpus among single systems, *i.e.*, non-ensemble system. In this section, we first describe the baseline systems and secondly we highlight on which subset of the data the baseline results are weaker, both for the LA and the PA task.

2.1. Baseline description

The LFCC (Linear Frequency Cepstral Coefficients) [5] are computed by taking the discrete cosine transform of the power audio spectrum. The first and second derivative coefficients are also concatenated to the original LFCC. The CQCC (Constant Q Cepstral Coefficients) [6] are computed by taking the cepstrum of the constant Q transform of the audio signal.

For each class (genuine speech and spoof speech), a Gaussian Mixture Model (GMM) [7] is trained on the corresponding features (LFCC or CQCC) using a pre-defined number of components (512 in the case of the baseline). For each test sample, a log-likelihood is computed for genuine and spoof classes and the so-called final score is the log-likelihood difference of these two classes log-likelihood. We reproduce the two baselines using the code provided by the organization (see Table 2 and 3) and find similar results as in [3].

2.2. LA: influence of the voice synthesis system

The per-category (of speech spoofing systems) results on the development set of the two baselines are available in [8]. Both the EER and min t-DCF obtained with the baselines are already low and they are equal to 0 for some speech spoofing systems. However, some speech spoofing systems introduce more errors than other, for instance US_1 and VC_4. Therefore, one of our contributions described in the next Section consists in taking into account the different speech spoofing systems during the model training to better solve the LA task.

2.3. PA: influence of category's ranges

The detailed per-category range results on the development set of the two baselines are given in [8]. However, Table 4 sums up the most interesting results for the baseline. We can observe that both baselines have difficulties to correctly classify the audio sample when the reverberation time T60 is very low (condition "a" of R) or when the replay device quality is perfect (condition "A" of Q). Thus, we can deduct that the reverberation information and the effect due to the replay device quality are meaningful information for the replay-attack detection (considering that the baseline is good enough). Moreover, when both conditions are present the results are even worst (*e.g.*, 36.7% EER and 0.811 min t-DCF for the CQCC-GMM). We can also observe that the LFCC are better designed than the CQCC for classifying sample with small T60. Conversely, the CQCC perform better in the case of perfect replay devices. Therefore, our contributions described in the next Section consist in taking into account all these observations to better solve the PA task.

[5]Later referred to as † and ♣.

3. Contributions

3.1. Category-specific spoof GMMs

Our main contribution is a classification method that exploits the "category" information during the learning. Following the observation drawn in Section 2, N specific GMMs are trained on N different subsets of the spoof training corpus (the investigated subsets are described hereafter). The same learning routine as in the baseline is followed, *i.e.* 10 iterations of an EM algorithm. The overall log-likelihood is then computed at the classification step following:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{genuine}} - \log\left(\sum_{n\in\mathcal{E}} \exp(\mathcal{L}_n)\right), \qquad (1)$$

where $\mathcal{E}$ is the set of N subsets and $\mathcal{L}_n$ is the log-likelihood outputted by n-th spoof GMM models. This principle is used by the three following classifiers.

Multi-categories (LA task). One GMM is trained for each of the six types of "speech spoofing system" in $\mathcal{E} = \{SS_1, SS_2, SS_4, US_1, VC_1, VC_4\}$. These spoof GMMs are composed of 85 components and the genuine GMM of 512. Thus, the total number of GMM components is equivalent to the one in the baseline. This system is referred to as *multicat* in Table 2.

The 3-spoof-GMMs (PA task). One spoof GMM is trained for each three types of Q in $\mathcal{E} = \{\text{Perfect}, \text{High}, \text{Low}\}$, for a total of four $(3+1)$ GMMs composed of 512 components.

The 9-spoof-GMMs (PA task). One spoof GMM is trained for each (9) combinations of ranges in Q $= \{\text{Perfect}, \text{High}, \text{Low}\}$ and in R $= \{\text{50-200}, \text{200-600}, \text{60-1000}\}$, for a total of ten $(3\times3+1)$ GMMs composed of 512 components.

3.2. Dual-LFCC features.

Our second contribution is a new type of feature called "Dual-LFCC". The "Dual-LFCC" feature vector is a concatenation of the LFCC features computed on the dereverberated signal and on the residual reverberation of the dereverberation processing. The dereverberation is achieved using the Doire algorithm [9], the residual is obtained by subtracting the dereverberated signal to the original signal and the LFCC are computed with the same parameters as in the baseline. This contribution is targeting the PA task and in particular the subset of examples where R is in the "a" range, *i.e.* examples where the reverberation provides little information to the classifier. The design intention of the Dual-LFCC is to highlight this reverberation information to reduce the complexity of the classification task. Similar features such as residual-LFCC, dereverb-LFCC and CQCC counterparts have been tested but with less success than the Dual-LFCC (the numerical results are not reported in this paper).

3.3. The 9-layers-DNN.

This contribution focuses on the use of a classical Deep Neural Network (DNN). The aim is to evaluate in which proportion the GMM is lacking modelling complexity. The considered DNN is a 9-layers network built in an autoencoder way where the number of neurons per layers is $(250, 200, 150, 100, 150, 200, 250, 90, 2)$. The end layer is a softmax layer with 2 neurons, which can be interpreted as the two posterior probabilities of the genuine and spoof class. The learning is achieved with 10 epochs using Adam optimizer and binary cross-entropy loss. Batch normalization is used between each layer.

Table 2: *Results (EER in % and min t-DCF) on the development and evaluation set for the LA task.*

Set	Classifier	Feature	EER (%)	min t-DCF
Dev	Baseline	LFCC	2.71	0.0663
Dev	Baseline	CQCC	0.43	0.0123
Dev	Multicat	CQCC	**0.16**	**0.0032**
Eval	Multicat	CQCC	7.95	0.2298

4. LA experiments

The overall results on the LA task are available in Table 2. The CQCC-GMM baseline results were already high on the development set, however our proposed system, *i.e.* the multi-categories GMM with the CQCC features, demonstrates improvements both in terms of EER and min t-DCF. In particular, the multi-categories classifier shows higher performances in the categories where the worst results were observed (see details in [8]). However, while in most categories our system show 0% EER on the development dataset, we infer from the results on the evaluation set the training data have been over-fitted and that this system does not generalize. We did not implement any fusion strategies for this task due to the effectiveness of our single system on the development set. For the record, the submitted primary system to the ASVSpoof challenge is ranked 32 over 50 submissions which is equivalent to the baselines rankings.

5. PA experiments

In this Section, the different experiments targeting the PA task are presented. A total of twelve systems are trained to address the PA task, twelve corresponding to all combinations of the three different types of feature (LFCC, CQCC and Dual-LFCC) and the four classifiers (Baseline, 3-spoof-GMMs, 9-spoof-GMMs and 9-layers-DNN) (see Table 3 and Section 5.1). The category range specific results of those twelve systems are then analyzed (see Table 4 and Section 5.2) with the target of mining useful information on the specificity of each system. Based on this information and on the results of the exhaustive list of all possible late fusion [10] of those twelve systems (see Section 5.3), the ASVSpoof 2019 challenge submission is then designed and discussed (see Table 5 and Section 5.4).

5.1. Overall results analysis

The results on the development set for each combination of classifiers / features are available in Table 3a for the EER and in Table 3b for the min t-DCF. For each kind of feature (including Dual-LFCC), both multi-categories settings, *i.e.*, 3- and 9-spoof-GMMs, improve the EER compared to the baselines but the min t-DCF remains approximately the same. As for the LA task, it seems that this setting succeed in take into account the variability of the problematic categories (in particular the reverberation time and the replay device quality). We can also notice that the DNN and the Dual-LFCC do not improve the overall results. Furthermore, their combination leads to even worse results. One explication could be that DNN is well suited for low level features, *i.e.* raw features like unprocessed spectrum, while Dual-LFCC are high level features.

The symbols ($\dagger, \clubsuit, \ddagger, *, \diamondsuit, \bigstar, 6, \heartsuit, 8, \sharp, \spadesuit, \natural$) present in Table 3 represent the twelve systems. They are used in Table 4 to refer to the different ensemble system described in the coming sections.

Table 3: *Results on the development set for the PA task.*

(a) *EER in %*

↓ Classifier ↓	LFCC	CQCC	Dual-LFCC
Baseline	$10.77^{\dagger}$	$9.92^{\clubsuit}$	$10.61^{\ddagger}$
3-spoof-GMMs	10.37^{*}	$\mathbf{9.02}^{\diamondsuit}$	$10.26^{\bigstar}$
9-spoof-GMMs	10.14^{6}	$9.13^{\heartsuit}$	9.56^{8}
9-layers DNN	$13.07^{\sharp}$	$12.78^{\spadesuit}$	$14.11^{\natural}$

(b) *min t-DCF*

↓ Classifier ↓	LFCC	CQCC	Dual-LFCC
Baseline	$0.226^{\dagger}$	$0.193^{\clubsuit}$	$0.255^{\ddagger}$
3-spoof-GMMs	0.234^{*}	$\mathbf{0.191}^{\diamondsuit}$	$0.257^{\bigstar}$
9-spoof-GMMs	0.229^{6}	$\mathbf{0.191}^{\heartsuit}$	0.232^{8}
9-layers DNN	$0.295^{\sharp}$	$0.251^{\spadesuit}$	$0.302^{\natural}$

5.2. Per-category range results analysis

The per-category range results of the twelve evaluated systems are available in Table 4. The results are focused on the range "a" of R and "A" of Q that were identify as the problematic ranges for the PA task. The first two lines correspond to the baselines. Most of the proposed systems improve the results on these specific category ranges (in EER or in min t-DCF). The best results are highlighted in bold numbers.

Regarding the range "a" of R, the LFCC baseline ($\dagger$) was originally better performing than the CQCC ($\clubsuit$). The new LFCC systems ($*$, 6, $\sharp$) are unsurprisingly among the best, in particular the ($*$) system has the best EER. However, the DNN CQCC ($\spadesuit$) provides the best min t-DCF for this range. These two last systems are part of our primary submission to the challenge as they bring the best results for the range "a" of R. Regarding the range "A" of Q, the CQCC baseline ($\clubsuit$) was originally better performing than the LFCC ($\dagger$). CQCC multi-categories systems ($\diamondsuit$, $\heartsuit$) are predictably the best systems in terms of min t-DCF for this range. It can be noticed that our primary submission to the challenge uses ($\heartsuit$).

It is interesting to note that we design the Dual-LFCC to solve the reverberation time issue. However, in practice the results are good on this case but not the best while conversely the system 8 (respectively $\bigstar$) is the best (respectively the third) system in terms of EER when the replay device quality is perfect. We have not yet found an explanation to this phenomenon and Dual-LFCC are not used in our submission anyway.

Furthermore, the systems identified in Table 4 significantly improve the overall results (as part of ensemble systems) while their individual overall results on the development set (displayed in Table 3) are comparable with the baselines. This tends to validate our strategy against problematic ranges.

5.3. Fusion experiments

Previous experiments [10] showed that late fusion of the results tremendously improved the results on this task. So, we decide to evaluate every possible combinations of the twelve systems. All combinations result in $\sum_{k=2}^{12} \binom{12}{k} = 4083$ different ensemble systems. For each ensemble system, a fusion function is trained using the computed scores on the development dataset of the corresponding systems. We use the Bosaris toolkit [11] for this training. Those fusion functions are then used to compute

Table 4: *Category range specific results. PA task.*

Range→	a of R		A of Q	
↓ System ↓	EER	t-DCF	EER	t-DCF
†	16.22	0.290	23.13	0.512
♣	17.88	0.332	21.33	0.458
‡	15.54	0.314	20.29	0.470
∗	**11.98**	0.252	20.79	0.506
♢	17.17	0.355	18.35	0.420
★	12.94	0.280	18.31	0.450
6	12.54	0.271	19.83	0.487
♡	17.72	0.362	17.94	**0.413**
8	12.99	0.298	**17.37**	0.450
♯	14.72	0.287	24.91	0.656
♠	13.72	**0.237**	26.16	0.610
♮	16.55	0.311	27.68	0.689

Table 5: *Ensemble systems for the PA task.*

Metric		EER (%)		min t-DCF	
Set		Dev	Eval	Dev	Eval
Baseline LFCC (†)		10.77	13.54	0.226	0.301
Baseline CQCC (♣)		9.92	11.04	0.193	0.245
Submission	Single (♡)	9.13	11.91	0.191	0.291
	Primary (∗♣♡♠ = $\mathcal{P}$)	7.37	10.53	**0.140**	0.249
	Contrastive 1 (∗♢)	7.94	**9.59**	0.156	**0.225**
	Contrastive 2 ($\mathcal{P}$†★)	**7.31**	10.53	0.145	0.255
Best EER ($\mathcal{P}$♢†‡♮♯)		**6.80**	-	0.142	-
Best min t-DCF ($\mathcal{P}$♯)		7.22	-	**0.135**	-
12 systems		7.07	-	0.148	-
Best without DNN (♣♡6★‡)		8.13	-	0.156	-
Best without DNN (♣ ∗ ♢‡)		7.85	-	0.158	-
Baseline fusion (†♣)		9.22	-	0.178	-

the scores of the ensemble systems on both the development and evaluation datasets. We then analyze the results on the development set of those 4083 ensemble systems both in terms of EER and min t-DCF in order to design our submission to the challenge. The systems identified here significantly improve the overall results (as part of ensemble systems) while their individual overall results showed in Table 3 are comparable with the baselines.

5.4. ASVSpoof challenge submission

Our submission to the challenge is based on the results of single and ensemble systems on the development set while the challenge evaluate on a different set of data. The submitted single system is the 9-spoof-GMMs CQCC (♡) system which corresponds to the best system with respect to the min t-DCF. The primary system is composed of four systems (∗♣♡♠ = $\mathcal{P}$) and is built following the observation drawn in Section 5.2. It is also the best ensemble system in terms of min t-DCF with four or less systems. The Contrastive 1 is composed of two systems (∗♢). It is the best ensemble system both in terms of EER and t-DCF with only two systems. The Contrastive 2 is composed of six systems ($\mathcal{P}$†★). It is the best ensemble system with six or more systems. The results of these ensemble systems are available in Table 5 both for the development set and the evaluation set.

While the three submitted ensemble systems demonstrate major improvements on the development set (*e.g.*, 0.193→0.140), their results on the evaluation set are disappointingly comparable to the baselines. The primary system is ranked 35 over 52 submissions which is equivalent to the baselines rankings. Conversely, the Contrastive 1 system shows a slight improvement compared to the baselines (0.245→0.225). As explained in the evaluation plan [3], speakers and room impulse responses changed from the development set to the evaluation set. The over-fitting of our method on the PA task can be explained more likely by over-fitting room impulse responses than speaker as ranges of reverberation time T60 (R) are used in the training procedure whereas speaker labels are not.

In addition to our submission results, Table 5 also display for the results of various other typical ensemble systems for the development set only as the ground truth of the evaluation set is not disclosed to the participants yet. This results are not commented further but give the reader a broader view of ensemble systems.

5.5. Discussion and perspectives

First, we noticed that among the 4083 fusion combinations, the ensemble systems comprising the DNN are among the best, whereas the DNN alone does not improve the results. The considered structure is simple yet effective in this case. One explanation could be that the DNN explores a larger space than the GMM and in doing so adds an additional generalization power to the overall ensemble system.

Second, our preliminary experiments show that the baselines has more difficulties to correctly classify the sound examples when the range of the R category (T60) was the smallest, *i.e.*, "50-200"ms. To this end, we considered the Dual-LFCC, which corresponds to the concatenation of the dereverberated and residual signals. Whereas this feature does not improve the results by itself, it allows to build an ensemble system which has the lower EER among the other systems. As the dereveberation [9] uses external data (forbidden by the organization), we did not submit to the challenge any system including the Dual-LFCC. However, the external data could be replaced by the challenge data but we were lack time to implement it. The release of the labels of the evaluation set will enable us future research perspectives. In particular, we will be able to evaluate the generalization capability of the Dual-LFCC.

6. Conclusions

This paper presents our contributions to the 2019 ASVSpoof Challenge (LA and PA tasks). Our first contribution is a training procedure that exploits the labelling of the attack conditions in the training set. In particular for the replay attack scenario, we identify low reverberation time and perfect microphone to be problematic attack conditions and allocate more modelling power to these training subsets. Our second contribution is a new feature called Dual-LFCC that target the low reverberation time attack condition. We then conduct extensive experiments on single (resp. ensemble) systems, which lead to major improvements of the baseline on this particular attack conditions (resp. on all types of attack). However, we suspect that our training procedure has over-fitted the development set as similar improvements are not observed on the evaluation set. This still need to be confirmed for the Dual-LFCC, once the evaluation set will be released.

7. 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, pp. 130–153, 2015.

[2] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[3] ASVspoof consortium, "Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," p. 19, 2019.

[4] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Odyssey 2018 The Speaker and Language Recognition Workshop*. ISCA, 2018, pp. 312–319.

[5] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH 2015*, 2015, pp. 2087–2091.

[6] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: a spoofing countermeasure for automatic speaker verification," *Computer, Speech and Language*, vol. 45, pp. 516–535, 2017.

[7] G. McLachlan and D. Peel, *Finite Mixture Models*. Wiley, 2000.

[8] N. Souviraà-Labastie, M. Baelde, T. Malet, and R. Greff, "Impact des conditions d'attaques sur les contre-mesures pour la reconnaissance du locuteur," in *Gretsi 2019 (submitted)*, 2019, pp. 1–4.

[9] C. S. J. Doire, M. Brookes, P. A. Naylor, C. M. Hicks, D. Betts, M. A. Dmour, and S. H. Jensen, "Single-Channel Online Enhancement of Speech Corrupted by Reverberation and Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 572–587, Mar. 2017.

[10] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion," in *INTERSPEECH 2018*. ISCA, 2018, pp. 77–81.

[11] N. Brümmer and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classier Score Processing," *Documentation of BOSARIS toolkit*, p. 24, 2011.