INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Cross-database replay detection in terminal-dependent speaker verification

*Xingliang Cheng, Mingxing Xu and Thomas Fang Zheng**

Center for Speech and Language Technologies,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing, China
fzheng@tsinghua.edu.cn

Abstract

The vulnerability of automatic speaker verification (ASV) systems against replay attacks becomes a severe problem. Although various methods have been proposed for replay detection, the generalization capability is still limited. For instance, a detection model trained on one database may fully fail when tested on another database. In this paper, we adopt the one-class learning technology to address the cross-database problem. Different from conventional two-class models that discriminate genuine speeches from replay attacks, the one-class model focuses on the within-class variance of genuine speeches, which is naturally robust to unseen attacks. In this study, we choose the Gaussian mixture model (GMM) as the one-class model and design two utterance-level features which reduce the uncertainties of genuine class while still be distinguishable from non-genuine class. Experiments conducted on three public replay datasets show that, compared to the state-of-the-art methods, the proposed method demonstrates promising generalization capability under cross-database scenarios.

Index Terms: replay detection, one-class learning, automatic speaker verification

1. Introduction

With the development of automatic speaker verification (ASV) technology, it is increasingly deployed in many applications. In the meantime, the vulnerability of ASV systems under spoofing attacks becomes a severe problem. There are four kinds of spoofing attacks [1] including impersonation, speech synthesis (SS), voice conversion (VC) and replay. Compared to SS and VC, replay attacks require neither specific expertise nor sophisticated equipment (a mobile phone is enough), showing a high threat to ASV systems [2]. Countermeasures therefore must be developed to detect replay attacks.

In principle, there are two kinds of approaches for replay detection: feature-based approaches and model-based approaches. For feature-based approaches, researchers have proposed a series of well-designed features to discover the distortion during the replay process, such as the constant-Q cepstral coefficient (CQCC) [3], linear prediction cepstral coefficient (LPCC) [4], and inverse Mel frequency cepstral coefficient (IMFCC) [5]. For the model approach, both statistical model and deep neural model are proposed to classify the genuine speech and replayed speech [6, 7, 8].

Although these methods perform well on a specific database, the generalization capability is still limited when it is deployed under unseen replay conditions. A previous study [9] shows that performance was significantly degraded when the training and test data come from different databases. To improve the generalization capability, Wang *et al.* [10] proposed a domain adversarial training technique to improve the cross-database robustness. The feature extractor was designed to extract spoofing cues and remove domain differences by the adversarial training technique. It prevents models from overfitting on the bias presented on the training set. However, it is still limited to deal with unknown attacks since it cannot discover the fake cues absent in the training set.

In this paper, we argue that the principal difficulty of this generalization problem attributes to the large number of attack compositions. For instance, if there are 10 recording devices and 10 playback devices, they produce 100 attack compositions. Furthermore, different playback and recording devices have different physical attributes, leading to unpredictable replay data distribution. Most existing approaches to replay detection are based on a two-class model, aiming to detect whether a speech is genuine or replayed. Due to the large number of recording-replay compositions, a two-class model is naturally overfitted on the training data and cannot deal with other unseen compositions.

To overcome this problem, we transfer the two-class (genuine/replay) classification task into a one-class (genuine) detection task. We argue that if the genuine data distribution is well modeled, it is robust to deal with various kinds of unseen replay attacks. To ensure the generalizability of the one-class model, the variability of genuine speeches needs to be reduced. Two methods, namely the long-term averaged spectrum estimation and the phone-attentive averaged spectrum estimation, are designed to reduce the short-term variability (such as speech contents). Then, the spectrum difference between the enrollment speeches and test genuine speeches is obtained to reduce the long-term variability (such as speaker traits and recording devices). Finally, the residual variability of genuine speeches is modeled by a one-class GMM. Experiments were conducted on terminal-dependent ASV tasks where the recording device of each speaker is fixed. Results show that the proposed method has strong generalization capability under cross-database scenarios and works well on real-life replay databases.

The rest of this paper is organized as follows. Section 2 reviews related works on replay countermeasures. Section 3 describes the proposed method. Experiments and results are reported in Section 4. We conclude the work in Section 5.

2. Related Work

One-class model for anti-spoofing. Researchers have explored one-class approaches on SS and VC attacks. Alegre *et al.* [11] used the one-class support vector machine (OCSVM) to estimate the boundary of genuine samples in the local binary patterns (LBP) feature space. Experiments shown that OCSVM was more robust than the normal support vector machine (SVM)

* Corresponding Author

http://dx.doi.org/10.21437/Interspeech.2021-960

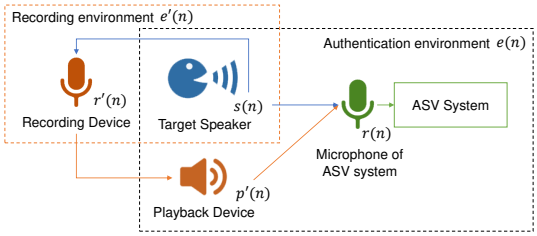

Figure 1: *The process of genuine access and replay attack.*

model since the latter failed to detect unseen attacks. Villaba *et al.* [12] proposed a scheme using DNN models to automatically learn discriminative features for OCSVM training. In this paper, we adopt the one-class method for replay attack detection with specially designed utterance-level features.

Speaker-specific replay detection. Suthokumar *et al.* [13] utilized the enrollment data of ASV systems to boost the performance of replay detection. A strong prior assumption is that the enrollment phase is often human-supervised, and the enrollment utterances are naturally genuine. Based on this assumption, all genuine speeches were firstly modeled by a Gaussian mixture model (GMM), which is also named as universal background model (UBM). Then, based on the individual enrollment data, a speaker-specific GMM model was adapted from the genuine UBM by the maximum a posteriori (MAP) estimation. This paper adopts this assumption and designs speaker-irrelevant features to improve the generalizability in replay detection tasks.

3. Method

In this section, we will describe the proposed one-class approach for replay detection. Firstly, two utterance-level features, namely, the long-term averaged spectrum based residual variability and the phone-attentive averaged spectrum based residual variability, are extracted to reduce the influence of speech contents, speaker traits, and devices. Then, the residual variability of genuine speeches is modeled by a one-class GMM for replay detection.

3.1. Residual variability feature

As shown in Figure 1, the overall process of replay attacks involves recorders, loudspeakers, and environments. Considering the high quality of microphones and loudspeakers, we can treat the replay process as a linear time-invariant (LTI) system. Under this assumption, a genuine speech can be defined as:

$$X_g(k,l) = S(k,l)E(k)R(k), \tag{1}$$

where $X_g(k,l)$ and $S(k,l)$ are the spectrogram of a genuine signal and the corresponding speech signal around the mouth of the target speaker, respectively; k is the index of frequency bins, and l is the index of frames; $E(k)$ and $R(k)$ are the frequency responses of authentication environments and the microphone of ASV systems, respectively; We ignore the effect of additive noises in this definition.

Also, a replay signal can be defined as:

$$X_r(k,l) = S(k,l)E'(k)R'(k)P'(k)E(k)R(k) \tag{2}$$

where $X_r(k,l)$ is the spectrum of a replay signal; $E'(k)$, $R'(k)$, and $P'(k)$ are the frequency response of recording environments, recording devices, and playback devices of the replay process, respectively.

As discussed in Section 1, due to the huge number of types of recording and playback devices, it is impractical to model all replay attacks X_r based on very limited attack compositions. For example, in the ASVspoof 2017 database, there are only 3 compositions in the training set and 10 compositions in the development set. However, we can rewrite the equation (2) as

$$\begin{aligned} log(|X_r(k,l)|) = log(|X_g(k,l)|) \\ + log(|E'(k)R'(k)P'(k)|) \end{aligned}. \tag{3}$$

The equation (3) shows that the replay process shifts the genuine speech away in the log magnitude domain. If the distribution of genuine speeches can be modeled well, we can easily detect those shifted replay samples. However, the variability of genuine speech makes it difficult to model. Actually, we do not need to model all those variabilities since our goal is to detect non-genuine speeches, not to generate speeches. Thus, we can compress the feature space of genuine speeches by eliminating the variability caused by various factors, such as speech contents, speaker traits, recording devices, etc., and then model the residual variability (RV) by one-class learning.

3.1.1. Long-term averaged spectrum based residual variability

Genuine speeches are composed of short-term factors (such as speech contents) and long-term factors (such as speaker traits). The short-term factors can be filtered away by averaging the log-magnitude spectrogram along the time axis, which is also known as the long-term averaged spectrum (LTAS) [14]:

$$\text{LTAS}_x(k) = \frac{1}{L}\sum_{l=1}^{L} log(|X(k,l)|), \tag{4}$$

where $X(k,l)$ is the spectrogram of a signal $x(n)$, k is the index of frequency bins, l is the index of frames, and L is the total frame number of the signal $x(n)$.

Further, since both the speaker and recording device are the same in terminal-dependent ASV tasks, the long-term factors should be similar between the enrollment and test data. Thus, those long-term factors can also be eliminated as:

$$r^{\text{LTAS}}(k) = \text{LTAS}_x(k) - \text{LTAS}_{enroll}(k), \tag{5}$$

where $enroll$ is the concatenation of all enrollment speeches of the claimed speaker[1] , the $r^{\text{LTAS}}(k)$ is the designed feature that contains the residual variability of genuine speeches, which is named as LTAS-RV in short.

3.1.2. Phone-attentive averaged spectrum based residual variability

If the contents of enrollment and test data are the same, then the variability of speech contents could be reduced by averaging. Usually, however, they are not the same. Thus, there will be some distortion caused by the mismatch of speech contents. To further reduce the variability of speech contents, we estimate the phone-attentive averaged spectrum (PAAS), which is defined as:

$$\text{PAAS}_x(p,k) = \frac{1}{L}\sum_{l=1}^{L} P(p,l)log(|X(k,l)|), \tag{6}$$

where $P(p,l)$ is the posterior of phone p at frame l.

[1] In the training phase, for each speaker, a subset of his/her genuine speeches is selected as the enrollment data.

Table 1: *Distribution of the repartitioned databases.*

Database	Subset	# Spk.	# Genuine	# Spoof
BTAS-PA	train	14	1,685	2,800
	dev	14	1,649	2,800
	eval-enroll	16	80	0
	eval-test		1,763	4,800
AS17	train	10	1,507	1,507
	dev	8	760	950
	eval-enroll	17	170	0
	eval-test		1,058	12,008
AS19Real	eval-enroll	26	130	0
	eval-test		260	1,860
THCHS-30	train	57	13,369	0

In this way, the content difference between the enrollment and test data could be eliminated by a phone-attentive subtraction of PAAS:

$$r^{\text{PAAS}}(k) = \sum_{p} w_p(\text{PSAS}_x(p,k) - \text{PSAS}_{enroll}(p,k)), \quad (7)$$

where $enroll$ is the concatenation of all enrollment speeches of the claimed speaker, w_p is the prior weight of phone p, which is defined as:

$$w_p = \begin{cases} c_p/C & \text{if } c_p > \theta \\ 0 & \text{else} \end{cases}, \quad (8)$$

where $c_p = \sum_l P(p,l)$ is the soft count of phone p over both enrollment speeches and the test speech, $C = \sum_p c_p$ is the normalization term, and θ is the threshold to exclude low-frequency phones. The $r^{\text{PAAS}}(k)$ is the designed feature which contains the residual variability of genuine speeches, and is named as PAAS-RV in short.

3.2. One-class classification model

The Gaussian mixture model (GMM) is used to perform the one-class classification. The distribution of the genuine class is modeled as:

$$P(x|\theta) = \sum_{i=1}^{K} w_i \mathcal{N}(x|\mu_i, \Sigma_i) \quad (9)$$

where w_i is the mixture weights, $\mathcal{N}(x|\mu_i, \Sigma_i)$ is the Gaussian function with mean vector μ_i and covariance matrix Σ_i, and K is the total number of components. The expectation-maximization (EM) algorithm [15] is used to estimate the parameters of GMM. The log probability of a given speech $log(P(x|\theta))$ is then used to decide whether it is genuine or not.

4. Experiments

4.1. Database

To evaluate the proposed method, the enrollment data for each speaker are necessary. The default partitions of publicly available replay detection databases are not suitable for this study. So we firstly repartitioned them based on their original partitions, as shown in Table 1. The BTAS 2016 database [16], ASVspoof 2017 version 2 (AS17) database [17] and ASVspoof 2019 real physical access (AS19Real) database [18] contain the real replayed samples. We discarded the ASVspoof 2019 physical access database [18] since the replayed samples in this dataset are simulated by algorithm instead of replayed in reality. A subset of BTAS that only contains replayed data was used, which we called the BTAS-PA. Besides, we selected the 'laptop' as the ASV system microphone, and all the genuine data which is recorded by the 'phone' was removed. We adopted a random subset of THCHS-30 as a genuine-only database for one-class training. For the AS19Real and BTAS-PA databases, the speaker-specific enrollment set contains 5 genuine utterances per speaker. Other databases contain 10 genuine utterances per speaker. The rest formed the speaker-specific test set. The speakers who do not have any spoofing data were removed from the AS17 dataset. All the data recorded by 'asv03' devices were removed from the AS19Real database since the speaker labels of those data are incorrect. For BTAS-PA and THCHS30 database, only the training set was used for training. For ASVspoof 2017 database, the training and development set were pooled for training. The partition on the AS17 evaluation set is the same as the one used in [13]. The detail of the partition can be found at `https://github.com/aqzlpm11/SpeakerSpecificPartition`.

4.2. Experimental setup

To estimate the residual variability of genuine speeches, we selected K genuine utterances per speaker as the enrollment data. In the BTAS-PA training set, K is 5; In other datasets, K is 10.

The spectrogram can be estimated by either fast Fourier transform (FFT) or constant Q transform (CQT), which is the ancestor of LFCC or CQCC before the discrete cosine transform (DCT). Since the DCT of the average of the log magnitude spectrum equivalent to the average of the cepstral coefficients, we extracted the LFCC and CQCC for residual variability extraction. In order to be more comparable with the baseline, all the configurations of LFCC/CQCC extraction were the same as the baseline except that we used only the static part of LFCC/CQCC. Finally, a 20-dimension LFCC-LTAS-RV and 30-dimension CQCC-LTAS-RV were extracted. A 128-mixture one-class GMM was trained by using the genuine samples only.

For PAAS-RV features, we used a model[2] pre-trained on the Librispeech dataset to predict the posterior of phones in BTAS-PA, AS17, and AS19Real databases. For the THCHS30, a model[3] pre-trained on Multi-CN databases was used. The threshold θ of PAAS method was set to 1.

4.3. Baseline system

As the baseline method, the popular CQCC and LFCC features were extracted. The configurations of LFCC and CQCC were identical to the ASVspoof 2019 challenge baseline [18]. For the LFCC feature, the hamming window size and its hop length were set to 20ms and 10ms, respectively. The number of linear filter banks was 20. Finally, a 60-dimension LFCC was formed by appending delta and delta-delta coefficients. For the CQCC feature, the maximum frequency f_{max} was 8KHz, and the minimum frequency $f_{min} = f_{max}/2^9$. Totally 96 bins per octave were extracted. In the resampling process, the number of uniform samples in the first octave was 16. Finally, a 90-dimension CQCC was formed by appending delta and delta-delta coefficients. No normalization technique was applied for both CQCC

[2] `http://www.kaldi-asr.org/models/m13`
[3] `http://www.kaldi-asr.org/models/m11`

Table 2: *EER(%) results of systems trained on various datasets. Systems 1 to 8 are two-class and systems 9 to 12 are one-class.*

ID	System	Train on AS17			Train on BTAS-PA			Train on THCHS-30		
	Test on:	**AS17**	BTAS-PA	AS19Real	**BTAS-PA**	AS17	AS19Real	BTAS-PA	AS17	AS19Real
1	LFCC + GMM [18]	30.44	50.31	43.37	**0.46**	47.05	34.24	/	/	/
2	CQCC + GMM [18]	22.76	52.47	50.30	8.17	47.47	49.95	/	/	/
3	FFT + ResNeWt [7]	15.60	9.98	30.38	3.58	25.32	16.55	/	/	/
4	CQT + ResNeWt [7]	12.21	23.37	18.86	2.10	32.61	24.62	/	/	/
5	FFT + LCNN [19]	9.00	11.75	42.69	5.49	19.20	51.93	/	/	/
6	CQT + LCNN [19]	9.08	21.22	30.00	5.85	29.96	28.48	/	/	/
7	LFCC + GMM + MAP [13]	20.52	15.89	39.57	0.96	20.60	21.98	/	/	/
8	CQCC + GMM + MAP [13]	10.68	13.56	43.02	4.94	19.56	38.86	/	/	/
9	LFCC-LTAS-RV + GMM	12.38	7.94	12.69	9.92	10.97	8.50	9.64	11.63	7.69
10	CQCC-LTAS-RV + GMM	8.61	**4.07**	6.14	3.92	8.79	6.52	5.95	8.41	4.57
11	LFCC-PAAS-RV + GMM	11.53	9.25	10.38	4.82	10.68	3.07	6.35	10.40	2.77
12	CQCC-PAAS-RV + GMM	**8.41**	9.99	**6.11**	3.69	**8.61**	**2.77**	**4.99**	**8.32**	**1.88**

and LFCC features. Two 512-mixture GMMs were utilized to model the genuine and spoof samples in the training set. During the MAP adaptation, the mean, variance, and weights of genuine GMM were adapted with a relative factor of 1. We also provided two DNN-based systems as the baseline. The input of the network was the log magnitude of the spectrogram which was calculated by FFT or CQT. For the FFT-based spectrogram, the frame length was 25ms, the hop length was 10ms and the Hanning window was used. For the CQT-based spectrogram was extracted with the 32ms-length frame hopping, Hanning window, 11 octaves, and 48 bins per octave. The mean and variance normalization (MVN) per utterance was applied on the log magnitude of the spectrogram. The detail of the ResNeWt and LCNN can be found in [7] and [19], respectively. Except for the MAP-based baseline models, other baseline models did not use no enrollment data in the testing phase.

4.4. Results and analyses

Table 2 shows the performance of different systems trained on various datasets. When systems were trained on the AS17 dataset, the performance of these two-class baseline systems (systems 1 to 8) was reasonable on the AS17 evaluation set. However, they cannot work well on BTAS-PA and AS19Real evaluation datasets. Similar results can be observed when systems were trained on the BTAS-PA dataset. Those results demonstrate the limited generalization capability of these methods under cross-database cases. Comparing with these methods, the proposed LTAS-RV/PAAS-RV feature with one-class GMM (systems 9 to 12) achieved consistently better performance, demonstrating the effectiveness of our methods. Moreover, the PAAS-RV methods (systems 11 and 12) outperform the LTAS-RV methods (systems 9 and 10) in most cases, which indicates that the reduction of the content influence can further boost performance.

We also trained the one-class model on the THCHS-30 dataset, which is a completely out-of-domain dataset. It can be observed that systems 9 to 12 still achieved a comparable performance, indicating that the proposed methods can be trained on an arbitrary speech datasets.

To further explore the reason, we use the t-SNE [20] method to visualize the distribution of the CQCC and the proposed CQCC-PAAS-RV features. As shown in Figure 2, in the CQCC feature space, different datasets are distributed differently. Thus, in the cross-database case, the training distribution and test distribution are mismatched, leading to the performance degradation. However, in the CQCC-PAAS-RV feature space, distributions of genuine samples are overlapped over different datasets. Thus, it is robust in the cross-database case. Meanwhile, replay samples are still distributed differently over datasets due to different replay compositions, which is why we should use the one-class model instead of the two-class model.

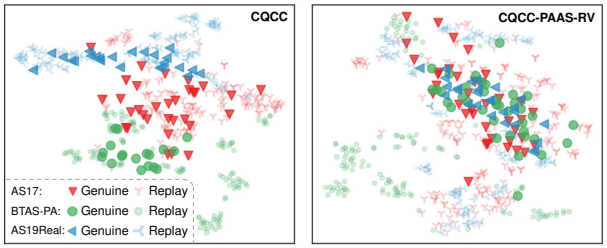

Figure 2: *The t-SNE visualization of the distribution of CQCC and CQCC-PAAS-RV features on three datasets.*

5. Conclusions

This paper proposed a one-class learning method for replay detection. The variability of genuine speech was be reduced by two methods, namely the long-term averaged spectrum-based method and the phone-attentive averaged spectrum-based method, with the help of the enrollment data in terminal-dependent ASV tasks. The residual variability of genuine speeches was modeled by a one-class GMM. Experiments conducted on several cross-database cases showed that the proposed method significantly outperformed the state-of-the-art methods, demonstrating the robustness and generalizability against unseen attacks.

6. Acknowledgements

This work was supported by the National Key Research and Development Program of China under the project No. 2017YFC0822204 and also the National Natural Science Foundation of China under the project No. 61633013.

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, Feb. 2015.

[2] M. Singh, J. Mishra, and D. Pati, "Replay attack: Its effect on GMM-UBM based text-independent speaker verification system," in *2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering (UPCON)*. Varanasi, India: IEEE, 2016, pp. 619–623.

[3] 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, Sep. 2017.

[4] M. Witkowski, S. Kacprzak, P. Żelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," in *Interspeech 2017*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 27–31.

[5] L. Li, Y. Chen, D. Wang, and T. F. Zheng, "A study on replay attack and anti-spoofing for automatic speaker verification," in *Interspeech 2017*, Stockholm, Sweden, 2017, pp. 92–96.

[6] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Interspeech 2019*. Graz: ISCA, Sep. 2019, pp. 1033–1037.

[7] X. Cheng, M. Xu, and T. F. Zheng, "Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China, Nov. 2019, pp. 540–545.

[8] B. Chettri, T. Kinnunen, and E. Benetos, "Deep generative variational autoencoding for replay spoof detection in automatic speaker verification," *Computer Speech & Language*, vol. 63, p. 101092, Sep. 2020.

[9] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," in *Interspeech 2016*, 2016, pp. 1705–1709.

[10] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Cross-domain replay spoofing attack detection using domain adversarial training," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 2938–2942.

[11] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. Arlington, VA, USA: IEEE, Sep. 2013, pp. 1–8.

[12] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Interspeech 2015*. Dresden, Germany: ISCA, Sep. 2015, pp. 2067–2071.

[13] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, E. Ambikairajah, and H. Li, "Use of claimed speaker models for replay detection," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Honolulu, HI, USA: IEEE, Nov. 2018, pp. 1038–1046.

[14] T. Kinnunen, V. Hautamaki, and P. Franti, "On the use of long-term average spectrum in automatic speaker recognition," in *ICASSP 2016 - 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kent Ridge, Singapore: IEEE, Dec. 2016, p. 9.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[16] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. G. Souza Mello, R. P. Velloso Violato, F. O. Simoes, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi, H. Dinkel, N. Chen, Y. Qian, D. Paul, G. Saha, and M. Sahidullah, "Overview of BTAS 2016 speaker anti-spoofing competition," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sep. 2016, pp. 1–6.

[17] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: Metadata analysis and baseline enhancements," in *Odyssey 2018 - The Speaker and Language Recognition Workshop*, 2018.

[18] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech 2019*. ISCA, 2019, pp. 1008–1012.

[19] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Dual-adversarial domain adaptation for generalized replay attack detection," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 1086–1090.

[20] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.