Interspeech 2022
18-22 September 2022, Incheon, Korea

Data Augmentation Using McAdams-Coefficient-Based Speaker Anonymization for Fake Audio Detection

Kai Li[1], *Sheng Li*[2*], *Xugang Lu*[2], *Masato Akagi*[1], *Meng Liu*[3], *Lin Zhang*[4], *Chang Zeng*[4], *Longbiao Wang*[3], *Jianwu Dang*[1], *Masashi Unoki*[1*]

[1]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[2]National Institute of Information and Communications Technology, Kyoto, Japan
[3] Tianjin University, Tianjin, China and [4]National Institute of Informatics, Tokyo, Japan

[1]{kai_li, akagi, jdang, unoki}@jaist.ac.jp, [2]{sheng.li, xugang.lu}@nict.go.jp,
[3]{liumeng2017, longbiao_wang}@tju.edu.cn, [4]{zhanglin, zengchang}@nii.ac.jp

Abstract

Fake audio detection (FAD) is a technique to distinguish synthetic speech from natural speech. In most FAD systems, removing irrelevant features from acoustic speech while keeping only robust discriminative features is essential. Intuitively, speaker information entangled in acoustic speech should be suppressed for the FAD task. Particularly in a deep neural network (DNN)-based FAD system, the learning system may learn speaker information from a training dataset and cannot generalize well on a testing dataset. In this paper, we propose to use the speaker anonymization (SA) technique to suppress speaker information from acoustic speech before inputting it into a DNN-based FAD system. We adopted the McAdams-coefficient-based SA (MC-SA) algorithm, and this is expected that the entangled speaker information will not be involved in the DNN-based FAD learning. Based on this idea, we implemented a light convolutional neural network bidirectional long short-term memory (LCNN-BLSTM)-based FAD system and conducted experiments on the Audio Deep Synthesis Detection Challenge (ADD2022) datasets. The results showed that removing the speaker information from acoustic speech improved the relative performance in the first track of ADD2022 by 17.66%.

Index Terms: fake audio detection, data augmentation, McAdams coefficients, speaker anonymization

1. Introduction

Recent advances in voice conversion (VC) [1] and text-to-speech (TTS) [2, 3, 4, 5] technology have made it possible to generate realistic and human-like speech for malicious purposes, such as spoofing and adversarial attacks. Studies on fake audio detection (FAD), which aims to distinguish spoof audios from real ones, are thus important and necessary to alleviate such threats. The current state-of-the-art FAD system is based on a deep neural network (DNN) model that is trained with a given dataset. In a DNN-based FAD system, the learning system may learn irrelevant information from the training dataset and could not generalize well on a testing dataset. Therefore, investigating how to remove irrelevant features from acoustic speech while keeping only robust discriminative features is essential for the FAD task.

Speaker information, an important biological characteristic for speaker recognition, is irrelevant (interference) information in the FAD task. Theoretically, speaker information can change when a different speaker's speech is selected as a reference in speech synthesis. However, speaker information cannot provide cues to distinguish synthetic speech from natural one; on the contrary, the appearance of speaker information can confuse the classification boundary between synthetic speech and natural speech. Therefore, we should suppress speaker information from acoustic speech for the FAD task.

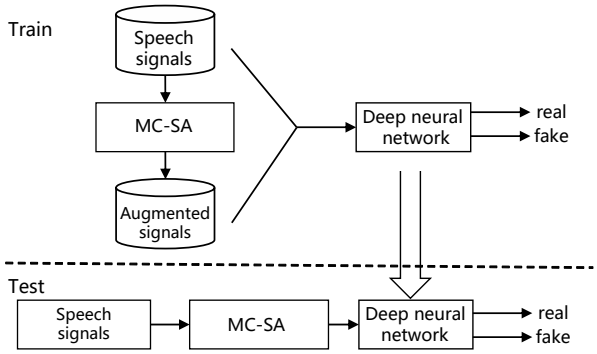
How to suppress speaker information entangled with other information in acoustic speech? One direct way is to use the speaker anonymization (SA) technique. The purpose of SA fits well with our task. According to studies in [6], a well-designed SA system should suppress speaker-specific information as much as possible, preserve intelligibility and naturalness, and protect voice distinctiveness. Anonymized speech should reduce the accuracy of automatic speaker verification (ASV) but should not affect the recognition performance in other missions such as automatic speech recognition (ASR). However, most SA algorithms are data-driven and require large computational resources, such as x-vector embeddings and neural waveform techniques [7, 8, 9]. Inspired by [10, 11], the McAdams coefficient (MC) based SA seems to be one of the best choices for our task. Based upon a simple contraction or expansion of pole locations derived using linear predictive coding (LPC), McAdams-coefficient-based SA (MC-SA) requires no training data and is comparatively straightforward and efficient. Therefore, the MC-SA algorithm is selected to suppress speaker information for the FAD task.

The MC-SA can be regarded as a new type of data augmentation (DA) method by which a global transform function derived from the MC-SA can be applied to generate acoustic speech for DNN-based FAD learning. DA methods generally have two goals. One is to increase the diversity of the training dataset and thereby prevent overfitting and improve robustness against out-of-domain data. This can be done, for example, by adding additional noise, using an acoustic codec [12], resampling the data, creating volume disturbance, or applying SpecAugment [13]. The other goal is to improve the accessibility of genuine and fake audio characteristics from the acoustic speech by suppressing interference information or revealing helpful information. For example, vocal tract length perturbation, which is derived from the vocal tract length normalization [14], has been used to reduce the effect of speaker information in ASR. The MC-SA-based DA method corresponds to the second goal. Therefore, different from most studies where MC-SA is used for privacy purposes, the MC-SA is used as a DA method for the FAD task in our study.

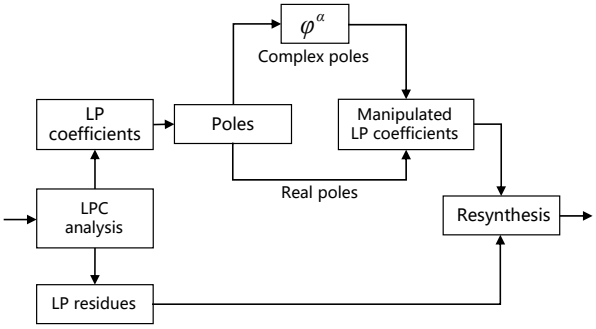
This paper proposes a novel DA method based on MC-SA

*Corresponding author.

10.21437/Interspeech.2022-10088



(a) Pipeline of proposed FAD system.



(b) Pipeline of MC-SA algorithm.

Figure 1: *Overview of FAD system based on proposed DA method.*

to suppress the interference of speaker information. Different MCs are used to generate anonymized speech for use in model training. It is supposed that the proposed DA method could conceal the speaker's information to some extent and reveal the acoustic features of fake audio without interference by speaker information. By training a DNN-based FAD system with the augmented speech data, improved performance is expected.

2. FAD system with proposed DA method

Figure 1 illustrates the working flow of a FAD system based on the proposed DA method. As shown in Fig. 1(a), the MC-SA algorithm, which is utilized as a DA block, is used to generate augmented (or perturbed) speech signals from the original ones. The augmented speech signals keep the same genuine/fake labels as the original ones. Then, the augmented speech signals are used to train a DNN model for FAD. The following two subsections describe the MC-SA algorithm and the design of the DNN model.

2.1. MC-SA algorithm

The MC-SA algorithm is illustrated in Fig. 1(b). The algorithm uses LPC analysis to obtain linear prediction (LP) coefficients as the vocal tract parameters. We can derive the position of poles from the LP coefficients. Then, the LP residuals, which corresponds to the glottal excitation, is calculated using inverse filtering. The glottal excitation or LP residuals is left for later resynthesis without any manipulation. Real-valued pole positions (with zero-valued imaginary terms) are left unmodified. In contrast, complex-valued poles (with non-zero imaginary terms) are shifted in accordance with the higher branch of Fig. 1(b), where the φ refers to the angle of poles with a non-zero imaginary part, and α is the MC [10]. Finally, we can resynthesize the anonymized speech signal using manipulated LP coefficients.

Angle φ and anonymized angle φ^{α} of poles correspond to different frequencies in the speech spectral envelope. A value of $\varphi = 1$ corresponds to a frequency of approximately 2.5 kHz. Seven spectral envelopes corresponding to different values of α ranging from 0.6 to 1.2 with increments of 0.1 are illustrated in Fig. 2. In Fig. 2, the solid black curve with $\alpha = 1.0$ is the original speech spectral envelope. As shown in Fig. 2, the manipulations in direction and scale are different depending on the distance from $\alpha = 1.0$. The spectral envelopes for $\alpha > 1.0$ are stretched along the frequency axis, whereas those for

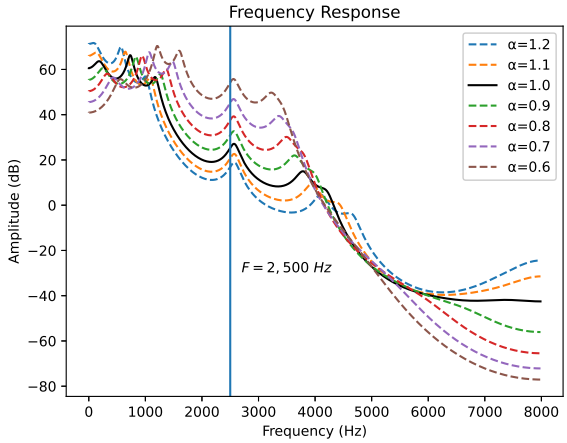


Figure 2: *Comparison of formant positions in spectral envelopes based on vowel /a/. Anonymized speech with different MCs ranging from 0.6 to 1.2 in increments of 0.1 are depicted.* $\alpha = 1.0$ *represents the original speech (solid black curve).*

$\alpha < 1.0$ are contracted. The scale of the shift depends on the distance from 2.5 kHz. The longer the distance, the larger the shift.

2.2. LCNN-BLSTM-based classifier

Some studies have shown that a shallow network as a back end is sufficient for downstream tasks [15], and this claim has been verified empirically in the anti-spoofing tasks [16]. Therefore, the DNN of the proposed FAD system was based on a previously used light convolution neural network (LCNN) [16] and was followed by two bi-directional recurrent layers using LSTM units (BLSTM), a global average pooling layer, and a fully connected output layer. The LCNN used max-feature-map (MFM) activation, which is based on the max-out activation function. The size of the BLSTM layers was equal to the dimensions of the LCNN's output. It is called LLGF network in [17, 18]. The detailed architecture of the LCNN-BLSTM model and the parameter settings for each layer are listed in Table 1.

The binary cross entropy (BCE) based objective function was used in model parameter optimization for FAD. The BCE

Table 1: *Architecture of LCNN-BLSTM-based deep classifier for FAD.*

Type	Kernel shape	Output shape	Params
Conv2d_0	[5, 5]	[64, 64, 253, 80]	1.664k
MaxFeatureMap2D_1	-	[64, 32, 253, 80]	-
MaxPool2d_2	-	[64, 32, 126, 40]	-
Conv2d_3	[1, 1]	[64, 64, 126, 40]	2.112k
MaxFeatureMap2D_4	-	[64, 32, 126, 40]	-
BatchNorm2d_5	-	[64, 32, 126, 40]	-
Conv2d_6	[3, 3]	[64, 96, 126, 40]	27.744k
MaxFeatureMap2D_7	-	[64, 48, 126, 40]	-
MaxPool2d_8	-	[64, 48, 63, 20]	-
BatchNorm2d_9	-	[64, 48, 63, 20]	-
Conv2d_10	[1, 1]	[64, 96, 63, 20]	4.704k
MaxFeatureMap2D_11	-	[64, 48, 63, 20]	-
BatchNorm2d_12	-	[64, 48, 63, 20]	-
Conv2d_13	[3, 3]	[64, 128, 63, 20]	55.424k
MaxFeatureMap2D_14	-	[64, 64, 63, 20]	-
MaxPool2d_15	-	[64, 64, 31, 10]	-
Conv2d_16	[1, 1]	[64, 128, 31, 10]	8.320k
MaxFeatureMap2D_17	-	[64, 64, 31, 10]	-
BatchNorm2d_18	-	[64, 64, 31, 10]	-
Conv2d_19	[3, 3]	[64, 64, 31, 10]	36.928k
MaxFeatureMap2D_20	-	[64, 32, 31, 10]	-
BatchNorm2d_21	-	[64, 32, 31, 10]	-
Conv2d_22	[1, 1]	[64, 64, 31, 10]	2.112k
MaxFeatureMap2D_23	-	[64, 32, 31, 10]	-
BatchNorm2d_224	-	[64, 32, 31, 10]	-
Conv2d_25	[3, 3]	[64, 64, 31, 10]	18.496k
MaxFeatureMap2D_26	-	[64, 32, 31, 10]	-
MaxPool2d_27	-	[64, 32, 15, 5]	-
Dropout_28	-	[64, 32, 15, 5]	-
BLSTM	-	[15, 64, 160]	154.880k
BLSTM	-	[15, 64, 160]	154.880k
FC	-	[64, 2]	322
Total	-	-	467.586k

is defined as:

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{N}[y_i \log P_\theta(\boldsymbol{x}_i) + (1-y_i)\log(1-P_\theta(\boldsymbol{x}_i))] \quad (1)$$

where N refer to the number of samples, θ denotes model parameters, the y_i and $P_\theta(\boldsymbol{x}_i)$ are the ground truth of the i-th training sample and its corresponding output probability from the model.

3. Experiments

3.1. Data and metrics

For evaluating the performance of our FAD system trained with the proposed DA method, we choose the data sets of the ongoing Audio Deep Synthesis Detection Challenge (ADD2022). The ADD2022 aims to further accelerate and foster research on detecting deep synthesis and manipulated audio. All tracks in ADD2022 share the same training and development datasets, while individual adaptation and validation datasets are released for fine-tuning and evaluation for each track. All these datasets involve more challenging attack situations in realistic scenarios compared with those in the ASVspoof2021 challenge [19]. Table 2 shows the statistical information for those three datasets.

Table 2: *Statistics for training, development, and adaptation datasets of track 1 of ADD2022 challenge. (Durations with three values denote min/mean/max.)*

	Genuine	Fake	Duration (sec.)
Training	3,012	24,072	0.86/3.15/60.01
Development	2,307	21,295	0.86/3.16/60.01
Adaptation	300	700	1.13/3.63/60.01

The utterance durations range from 1 to 60 s. The training and development datasets both contain genuine and fake utterances. The genuine utterances were obtained from a recently released large-scale high-fidelity multi-speaker Mandarin speech corpus called AISHELL-3 [20]. The fake utterances were generated using mainstream speech synthesis and VC systems. The speakers in AISHELL-3 are partitioned into two speaker-disjoint datasets for training and development, including about 50,000 utterances. The adaptation dataset for the first track includes 700 fake utterances generated using TTS and VC algorithms with various real-world noises and background music effects. The test dataset for the first track contains several unseen synthesized audios.

The performance is evaluated using equal error rate (EER) based metric, which is the same as used in the ADD2022 challenge.

3.2. Experimental setup

As MC-SA is used as a DA method, we need to find the best parameter setting for the FAD task. As we introduced in Section 2.1, the longer the distance from α to 1, the larger the spectral shift, and the more degradation on speech quality. Therefore, we set the minimum value of α to 0.6 to keep an acceptable speech quality. Also, the spectral envelopes for $\alpha > 1.0$ are stretched, which loses some information for FAD because some spectral envelopes will exceed the limitation of frequency band according to the Nyquist sampling theorem. Therefore, the maximum value of α was set into 1. We set the increment step of the MCs into 0.1.

The input to the network was an 80-dimensional log-Mel fbank feature. In the training stage, 4s segments are randomly selected from each raw waveform. Zero padding was applied to audios whose durations are shorter than 4s. The Mel spectrogram was extracted using the $MelSpectrogram$ module in the $torchaudio.transforms$ library [21]. Specifically, the size of fast Fourier transform, window length, and hop length in the short-time Fourier transform were set to 1024, 512, and 256, respectively. 64 utterances were grouped as one mini-batch to do feature extraction and fed into the DNN based FAD system; the number of training epochs was set to 50. The model that achieved the best results was compiled using an Adam optimizer with a learning rate value set at 0.0001.

4. Results and discussion

The effectiveness of the proposed DA method was evaluated by comparing its performance with that of the common DA method. The common DA method, including reverberation (R), diverse background noise (N), and music (M), was used to increase the diversity of the training dataset. In addition, spectral masking (SM) along the time and frequency axes were used further to improve the classifier's robustness [13]. The results

Table 3: *Preliminary results for FAD task based on the DNN FAD system trained based on SA and other compared DA methods in terms of EER.*

Training data	Augmentations	Results (EER %)	
		Adp. set	Val. set
Train+Dev.	R+N+M	3.83	33.40
Train+Dev.	R+N+M+SM	4.67	33.27
Train+Dev.	R+N+M+SM+MC (0.9)	3.62	32.22
Train+Dev.	R+N+M+SM+MC (0.8)	5.07	**30.32**
Train+Dev.	R+N+M+SM+MC (0.7)	6.38	31.08
Train+Dev.	R+N+M+SM+MC (0.6)	6.08	32.53
Train+Dev.+Adp.	R+N+M+SM	-	31.88
Train+Dev.+Adp.	R+N+M+SM+MC (0.8)	-	**26.25**

in terms of EER were separated into two parts in accordance with the difference in the training data and are listed in Table 3. The lower part shows the results for when the training, development (Dev.), and adaptation (Adp.) sets were used for training. Results were evaluated for Adp. set and online Val. set. The numbers in the parenthesis refer to the value of α.

The LCNN-BLSTM model combined with DA using additional reverberation, noise, and music achieved an EER of 33.40 on the validation dataset. Moreover, SM achieved a slight improvement over this (from 33.40% to 33.27%). As expected, the proposed augmentation method had a lower EER for all MCs settings. The best results were obtained when α was set to 0.8. Note that the best result on the validation set does not always correspond to the best adaptation result. This means that overfitting occurred in the training stage. The EERs with the proposed DA method were higher for the adaptation set but lower for the validation set. This means that the proposed method can alleviate overfitting and has an excellent ability to handle attacks of an unknown nature.

The training, development, and adaptation datasets were combined when conducting training to improve the performance on the validation set, which makes it difficult to identify the best epoch for performing the final validation. Therefore, the training was stopped empirically before convergence to avoid overfitting and obtain the best results. When the three datasets were used for training, the EER decreased from 31.88 to 26.25 %, which is a 17.66% improvement.

The separability can be visualized as a histogram of the classification probability. The distribution of classification probability for genuine and fake audios from the adaption dataset is depicted statistically in Fig. 3, where the horizontal axis refers to the classification probability of each utterance and the vertical axis refers to the number of audios. As we can see, the separation achieved with our proposed DA method (right histogram) is further enhanced. The number of audios appearing in a false class is lower. This suggests that the separation of our proposed augmentation method will be much better in the validation set.

5. CONCLUSION

In this paper, we proposed a novel DA method based on an SA algorithm for FAD. To suppress the interference of speaker information in the FAD task, we used the MC-SA algorithm to achieve a simple contraction of formant locations to conceal the speaker's identity. We also experimented with different MCs to identify the best non-linear frequency warping. Evaluations using the first track of the ADD2022 challenge demonstrated that,

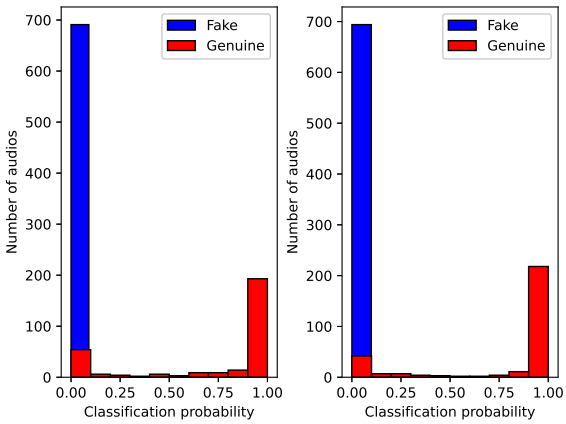


Figure 3: *Distributions of classification probability for genuine (red) and fake (blue) audios for the adaption dataset. The histogram on the left depicts model training results with common augmentation methods, including noise, reverberation, music, and spectral masking. The histogram on the right depicts model training results with the common augmentation method and proposed DA method.*

by suppressing speaker information, the proposed DA method can alleviate the overfitting problem to some extent and improve the ability to handle unknown attacks with advanced synthesis techniques. When combined with an LCNN-BLSTM classifier, the proposed DA method reduced the EER from 31.88% to 26.25% in online validation, which is a 17.66% improvement. Future work will focus on extending the proposed DA method for use in other speech and audio processing applications.

6. Acknowledgements

This work was supported by JSPS-NSFC Bilateral Joint Research Projects/Seminars (JSJSBP120197416), a Grant-in-Aid for Scientific Research (20H04207), a Grant-in-Aid for Early-Career Scientists (21K17837), the Fund for the Promotion of Joint International Research (Fostering Joint International Research (B))(20KK0233), the KDDI Foundation (Research Grant Program), NICT tenure-track funding, and NICT international funding. This work was performed when Kai Li was an intern at NICT.

7. References

[1] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[2] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *International Conference on Machine Learning*, pp. 4693–4702, 2018.

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, p. 3165–3174, 2019.

[4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu,

"Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[5] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for speech synthesis." *Proc. INTERSPEECH*, pp. 2027–2031, 2020.

[6] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J. F. Bonastre, "The voiceprivacy 2022 challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[7] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *arXiv preprint arXiv:1905.13561*, 2019.

[8] C. O. Mawalim, K. Galajit, J. Karnjana, S. Kidani, and M. Unoki, "Speaker anonymization by modifying fundamental frequency and x-vector singular value," *Computer Speech & Language*, vol. 73, 2022.

[9] J. M. Perero-Codosero, F. M. Espinoza-Cuadros, and L. A. Hernández-Gómez, "X-vector anonymization using autoencoders and adversarial training for preserving speech privacy," *Computer Speech & Language*, vol. 74, 2022.

[10] S. E. McAdams, *Spectral fusion, spectral parsing and the formation of auditory images*. Stanford university, 1984.

[11] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the mcadams coefficient," *arXiv preprint arXiv:2011.01130*, 2020.

[12] N. Hailu, I. Siegert, and A. Nürnberger, "Improving automatic speech recognition utilizing audio-codecs for data augmentation," *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–5, 2020.

[13] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "Specaugment on large scale datasets," *Proc. IEEE-ICASSP*, pp. 6879–6883, 2020.

[14] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[16] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks." *Proc. INTERSPEECH*, pp. 82–86, 2017.

[17] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," *arXiv preprint arXiv:2111.07725*, 2021.

[18] ——, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," *arXiv preprint arXiv:2103.11326*, 2021.

[19] J. Yamagishi, X. Wang, T. Massimiliano, S. Md, J. Patino, L. Xuechen, L. Kong Aik, K. Tomi, N. Evans, and D. Héctor, "ASVspoof2021: accelerating progress in spoofed and deep fake speech detection," in *Proc. ASVspoof 2021 Workshop*, 2021.

[20] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus," *Proc. INTERSPEECH*, pp. 2756–2760, 2021.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.