# Source Tracing: Detecting Voice Spoofing

Tinglong Zhu*, Xingming Wang*†, Xiaoyi Qin*†, Ming Li*†
* Data Science Research Center, Duke Kunshan University, Jiangsu, China
† School of Computer Science, Wuhan University, Wuhan, China
E-mail: tinglong.zhu@alumni.duke.edu xingming.wang@dukekunshan.edu.cn
xiaoyi.qin@dukekunshan.edu.cn ming.li369@duke.edu

***Abstract*—Recent anti-spoofing systems focus on spoofing detection, where the task is only to determine whether the test audio is fake. However, there are few studies putting attention to identifying the methods of generating fake speech. Common spoofing attack algorithms in the logical access (LA) scenario, such as voice conversion and speech synthesis, can be divided into several stages: input processing, conversion, waveform generation, etc. In this work, we propose a system for classifying different spoofing attributes, representing characteristics of different modules in the whole pipeline. Classifying attributes for the spoofing attack other than determining the whole spoofing pipeline can make the system more robust when encountering complex combinations of different modules at different stages. In addition, our system can also be used as an auxiliary system for anti-spoofing against unseen spoofing methods. The experiments are conducted on ASVspoof 2019 LA data set and the proposed method achieved a 20% relative improvement against conventional binary spoof detection methods.**

## I. Introduction

The performance of both text-to-speech (TTS) and voice conversion (VC) systems has improved significantly in the past few years with the significant development of deep learning [2] and advanced training strategies [3]. Many notable high-performance TTS and VC systems includes Tacotron systems [4], [5], Fast Speech systems [6], [7], VITS system [8], DelightfulTTS [9], etc. are proposed recently. This has exposed human users and ASV systems to increasingly serious potential attack threats and security concerns [10]. Therefore, building a trustworthy audio anti-spoofing system gradually attracts more and more attention. The most well-known fake audio detection challenge, the Automatic Speaker Verification and Spoofing Countermeasures (ASVspoof) challenge [11], [12], has been held since 2013 and focuses on building an audio spoofing countermeasure (CM) system for the ASV system. TTS/VS synthesized speech is often considered as the logical access (LA) attack. Generally, a CM system consists of a front-end feature extractor and a back-end classifier. The feature extractors typically extract handcrafted acoustic features based on the original waveform. Many acoustic features such as Constant Q Cepstral Coefficient (CQCC) [13], Group Delay Gram (GD Gram) [14], Joint Gram [15] and Inverted Mel-Frequency Cepstral Coefficients (IMFCC) [16] have been shown to be useful for audio anti-spoofing task. The back-end classifier usually identifies whether audio is a spoof or not based on the extracted features. More and more deep learning-based models and loss functions have been proposed to achieve better performance. In the latest ASVspoof2021 challenge, Tomilov et al. uses an LCNN-based [17] architecture and achieved impressive performance [18]. Despite there are many works on the CM system architecture, research on the problem of fake audio algorithm attributes analysis is relatively limited. Zhao et al. [19] uses the multi-task learning strategy to add the classification of known spoofing approaches to the existing CM framework and achieved noticeable performance improvement. However, this work only has an overall concept of detecting a set of systems and does not subdivide them according to different attributes at different levels. Therefore, this setup cannot handle unseen attacking scenarios. A similar solution was used by Borrelli et al. [20] to include a fake method detection module in a CM system using the multi-task learning approach. The unseen scenarios are considered as an open set classification problem. However, the approaches for generation are used for spoofing method classification.

In the field of deep forgery image and video detection, the problem of traceability about forgery algorithms has also attracted great attention in recent years. Jain et al. [21] uses six categories of face forgery algorithm labels as training targets for a forgery recognition system. In the testing phase, the authors achieved better generalization performance than a simple binary classification system by fusing all forgery algorithm categories and treating the system as a binary spoofing detection system.

TTS and VC systems can be divided into components such as speaker represent, waveform generator, and front-end models that convert text to a sequence of linguistic features [1]. An arbitrary spoofing system can be constructed by combining different components, making the CM system for LA access more challenging

In this work, we propose a framework for detecting spoofing attributes using multi-task learning to deal with the spoofing systems constructed by different combinations of TTS and VC modules. In other words, we want to use our framework to trace the attributes of an arbitrary speech synthesis or conversion system and determine what kind of algorithm is used during different stages. This could help to detecting unseen spoofing systems with one or more known attributes. For our work, we trace the following attributes of the complex LA spoofing systems:

- Conversion
- Speaker representation
- Waveform generator

We re-partition the training set and evaluation set of the

978-616-590-477-3 ©2022 APSIPA 

TABLE I
PARTIAL OF SUMMARY OF LA SPOOFING SYSTEMS [1]. * INDICATES NEURAL NETWORKS.

| | Input | Input processor | Duration | Conversion | Speaker representation | Waveform generator | Usage |
|---|---|---|---|---|---|---|---|
| A01 | Text | NLP | HMM | AR RNN* | VAE* | WaveNet* | Eval |
| A02 | Text | NLP | HMM | AR RNN* | VAE* | WORLD | Train |
| A03 | Text | NLP | FF* | FF* | One hot embed. | WORLD | Train |
| A04 | Text | NLP | - | CART | - | Waveform concat. | Train |
| A05 | Speech (human) | WORLD | - | VAE* | One hot embed. | WORLD | Eval |
| A06 | Speech (human) | LPCC/MFCC | - | GMM-UBM | - | Spectral filtering + OLA | Train |
| A07 | Text | NLP | RNN* | RNN* | One hot embed. | WORLD | Eval |
| A08 | Text | NLP | HMM | AR RNN* | One hot embed. | Neural source-filter* | Train |
| A09 | Text | NLP | RNN* | RNN* | One hot embed. | Vocaine | Train |
| A10 | Text | CNN+bi-RNN* | Attention* | AR RNN+CNN* | d-vector (RNN)* | WaveRNN* | Train |
| A11 | Text | CNN+bi-RNN* | Attention* | AR RNN+CNN* | d-vector (RNN)* | Griffin-Lim | Train |
| A12 | Text | NLP | RNN* | RNN* | One hot embed. | WaveNet* | Train |
| A13 | Speech (TTS) | WORLD | DTW | Moment matching* | - | Waveform filtering | Train |
| A14 | Speech (TTS) | ASR* | - | RNN* | - | STRAIGHT | Train |
| A15 | Speech (TTS) | ASR* | - | RNN* | - | WaveNet* | Train |
| A16 | Text | NLP | - | CART | - | Waveform concat. | Train |
| A17 | Speech (human) | WORLD | - | VAE* | One hot embed. | Waveform filtering | Train |
| A18 | Speech (human) | MFCC/i-vector | - | Linear | PLDA | MFCC vocoder | Train |
| A19 | Speech (human) | LPCC/MFCC | - | GMM-UBM | - | Spectral filtering+OLA | Train |

ASVspoof 2019 dataset [1] of the LA task for out experiments. Under our multi-task training strategy, we achieve 88.4% accuracy in identifying conversion algorithms, 51.5% accuracy in detecting speaker representation modules (acoustic models, speaker encoder, etc.) and 77.5% accuracy in identifying waveform generator by a single model. Moreover, our system can also be used as an auxiliary system for anti-spoofing detection to achieve better performance against unseen spoofing systems.

## II. PROPOSED METHOD

### A. Related Works

The existing work for tracing the spoofing methods are mostly be used to improve the performance of the anti-spoof detection system rather than classifying the synthesis methods. Li et al., [22] identity the spoofing algorithm as an additional task under a multi-task training framework to improve the system performance on top of the anti-spoof countermeasure task. But the generating methods of test audio are seen in the training set. Borrelli et al., [20] conduct research on detecting and classifying spoofing methods. However, neither of them have studied the attribute classification for spoofing attacks, [20] only classifies different spoofing methods based on the waveform generator, while Li et al., [22] detects the whole fake system without further subdivision. Adopting a set of attributes to describe the spoofing methods at different stages of the whole pipeline could enhance the robustness of the current spoofing method identification when detecting unseen spoofing methods. In order to improve the spoof detecting system's robustness towards those spoofing systems that are not directly included in the training set, but part of their modules are similar to the ones of other spoofing systems in the training set, we here propose a multi-task attribute classification training strategy. In this paper, we focus on classifying attributes of the spoofing systems.

### B. Multi-Label Classification

As shown in the training part of Fig. 1(b)), according to the generating pipeline of spoofing speech, there are three attributes that are most important: Conversion, Speaker Representation, and Waveform Generator.

*1) Conversion:* Here the conversion denotes the feature transformation modules. The goal is to convert the input feature to match the target speaker's voice.

*2) Speaker Representation:* Attackers can take advantage of speaker representation to imitate a target speaker's voice. This may include speakers registered in the ASV security systems. Typically speaker representation is a high-dimensional vector that contains the speaker's embedding or index in the training data, timbre, etc. With speaker representation, attackers are able to generate speech according to target speaker's characteristics using TTS or VC algorithms.

*3) Waveform Generator:* Waveform generator performs the conversion from acoustic features to the corresponding speech signals which are also called vocoders. The performance of waveform generator is highly correlated to the quality of the synthesized speech.

### C. Spoofing Attribute Classification

We propose a training strategy that different back-end classifiers share the same front-end model.

$$\mathbf{e}_i = Z(X_i) \tag{1}$$

Where $\mathbf{e}_i \in \mathbf{R}^d$, indicates the output vector extracted by front-end model $z(.)$ from $i_{th}$ audio. In this work, we define three spoofing attributes detection classifiers mentioned above: conversion, speaker representation, and waveform generator. The loss functions of classifiers are:

$$l_{conv} = L_{CE}(C_{conv}(e_i), y_i^{conv}) \quad (2)$$

$$l_{spk} = L_{CE}(C_{spk}(e_i), y_i^{spk}) \quad (3)$$

$$l_{wg} = L_{CE}(C_{wg}(e_i), y_i^{wg}) \quad (4)$$

Where $conv$ denotes conversion, $wg$ denotes waveform generator and $spk$ denotes speaker representation and $C_{conv}$,$C_{spk}$ and $C_{wg}$ represent the corresponding attribute classifiers. And $y_i^{class}$ denotes the predicted label for each corresponding attribute. We apply the Cross-Entropy loss as our loss function. The final loss is formulated as a weighted summation and in which $\lambda_i$ is the weight value for different attributes:

$$l_{total} = \lambda_1 l_{conv} + \lambda_2 l_{spk} + \lambda_3 l_{wg} \quad (5)$$

### D. Anti-spoofing Countermeasure

In addition to being able to trace the attribute of the spoofing system, the proposed method is also able to improve the performance of the anti-spoofing countermeasure system. Unlike most CM systems (Fig. 1(a)), which only has one classifier for determining whether the input speech utterance is spoofed, our system has three classifiers for spoofing detection. This enables us to combine the bona fide probability from each classifier to get the final probability of whether the it is spoofed or not (Fig. 1(b)). We adopt the cubic root of the product after multiplying three spoof probabilities from the classifiers as the final spoof score.

$$s_{spoof} = \sqrt[3]{p_{spoof_{conv}} \times p_{spoof_{wg}} \times p_{spoof_{spk}}} \quad (6)$$

## III. Experiment Setup

### A. Dataset Division

In our experiments, we use the ASVSpoof 2019 LA task's dataset [1], which has 121461 utterances in total. As shown in Table. I, the methods used in the original evaluation set and development set (A07-A19) are not fully covered in the original training set (A01-A06). Thus we reconstruct the training set and evaluation sets. To make sure the attributes of all the methods in the evaluation are covered in the training set and there is no speaker overlap between these sets, we first divided all speakers in the LA dataset to two parts, one as training speaker set and the other for evaluation speaker set. For our evaluation set ,we choose the utterances from our evaluation set speakers with label bona fide, A01, A05, and A07 to form our evaluation set. And we select utterances with label A02-A04, A06, A08-A19 from our training set speakers and these speakers' bona fide utterances to form our training set. Note that the division will inevitably leave some utterances unused in the experiment. In this setup, our training set contains 67 speakers and 79620 utterances, while the evaluation set has 11 speakers and 5832 utterances.

### B. Label Assignment

As mentioned above, we proposed tracing the spoofing system's attributes. There are three attribute labels to train our classifiers, which means each utterance has three labels. Table.II presents the detail of the label assignment. To make the model more generalized, we divided the waveform generator into NN-based and non-NN-based methods. In addition, we combined all RNN-related methods in conversion attributes to one label. We keep the original labeling with [1] for the speaker representation attribute.

TABLE II
LABELING FOR EACH ATTRIBUTE

| Attribute (Classifier) | Methods |
|---|---|
| Waveform generator | Nerual Network methods |
| | non Neural Network methods |
| | bona fide |
| Speaker represent | VAE |
| | One hot embed. |
| | d-vector (RNN) |
| | PLDA |
| | bona fide |
| Conversion | RNN related methods |
| | FF |
| | CART |
| | VAE |
| | GMM-UBM |
| | Moment matching |
| | Linear |
| | bona fide |

### C. Models

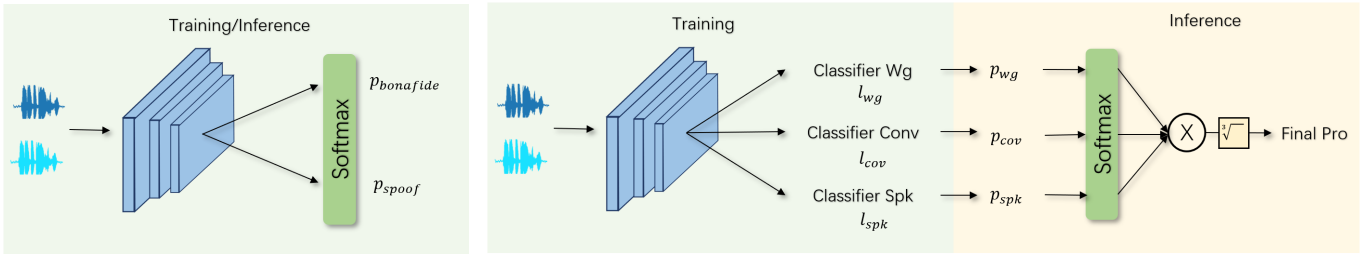
We validate our strategies using two models, where one is based on ResNet34 [23], and the other is based on the RawNet2 [24]. For the ResNet34 based model, we adopt the model structure of the ASV system proposed by Cai et al. [25]. We apply log-FBank algorithm with 80-dimension Mels for feature extraction for the ResNet34 systems. For the RawNet based model, we employ the SincNet [26] based RawNet2 system [24] in the experiment. We use the log-FBank features for the ResNet34 model, while for the RawNet2 model, we directly input the truncated/concatenated signals to the model. In a single experiment, all three classifiers receive the same inputs from the same front-end model.

## IV. Experiment Results

### A. Spoofing attribute classification

The performance overview of the ResNet34 and RawNet2 multi-task system are given in Table. III. We report each attribute's accuracy and clearly observe that the recognition accuracies of Conversion and Waveform generator attributes are over 80%. But on the other hand, the accuracy of speaker representation is only about 50%. This is due to the fact that the speaker representation is a latent vector built by the conversion model and not explicitly expressed on the signal.

Therefore, we further analyze and count the predicted results of speaker represent as shown in Fig. 2. The results presents




(a) (b)

Fig. 1. Comparison between conventional spoof detection system and multi-task learning based spoof detection system

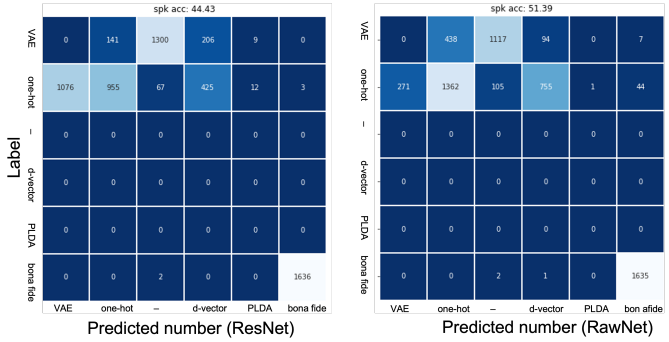



Fig. 2. Predicted result of different model under the speaker represent attribute

that although speaker represent attribute can not be correctly classified, the classifier still has a high accuracy in the discrimination of bonafide.

TABLE III
MULTI-TASK TRAINING ON ASVSPOOF 2019 LA DATASET

| Model | Conv. Acc(%) | Spk. Acc (%) | Wg. Acc (%) |
|---|---|---|---|
| ResNet34 | 86.5 | 44.43 | **84.47** |
| Rawnet 2 | **88.41** | **51.46** | 77.54 |

### *B. Spoofing detection*

As shown in Section. II-D, our method could not only track the fake audio's attribute but also implement spoofing detection. Since the part of evaluation data is considered as training set, we recontribute the test set using new evaluation data mentioned in III-A. In this experiment, we adopt the conventional spoof detection system, which implement the binary classifier for spoofing detection to determine whether the test audio is spoofed, as baseline system, named Binary. The results presents in Table. IV. Since training and evaluation set is small and the different training seed will slightly influence the results [27], all results reported in this paper are best result. As presented in Table. IV, our strategy can help improve the system performance by at least 20%. ResNet-based model even achieves about 80% relative improvement than baseline system. Since the original binary classification task (bona fide and spoof classification) is split to three attribute classifiers, the model space for the spoof detector has been increased, which may be a possible explanation for the performance improvement.

TABLE IV
PERFORMANCE COMPARISON BETWEEN CONVENTIONAL SPOOF DETECTION SYSTEM AND MULTI-TASK LEARNING BASED SPOOF DETECTION SYSTEM

| Model | Method | EER [%] |
|---|---|---|
| ResNet34 | Multi-task (proposed) | **0.012** |
| ResNet34 | Binary | 0.066 |
| RawNet2 | Multi-task (proposed) | **0.187** |
| RawNet2 | Binary | 0.241 |

## V. CONCLUSIONS

In this paper, we show that our multi-task training strategy can help the model achieve acceptable performance in attribute source tracing for logical access spoofed utterances. The experiment results shows that our strategy can not only help source tracing but also help to improve the spoof detecting systems by combining the bona fide probabilities from three attribute classifiers. In addition to whether the utterance is spoofed or not, our training strategy can help to extract extra information like how the spoof system is designed, what algorithm the spoof systems used to generate a waveform, etc., which can help to improve the robustness of the spoof detection system towards those spoofing systems that are not directly included in the training set, but part of their modules are similar to the ones of other spoofing systems in the training set. Moreover, we present a new strategy to improve the performance of the spoof detecting system in addition to improving the front-end model, feature extraction, etc. For the future exploration of the source tracing, more spoofed data generated by different combinations of spoofing algorithms are needed.

## ACKNOWLEDGMENT

This research is funded in part by the National Natural Science Foundation of China (62171207), Kunshan Municipal Government Research Funding. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

## REFERENCES


[1] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[2] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv:2106.15561*, 2021.

[3] F. Yue, Y. Deng, L. He, T. Ko, and Y. Zhang, "Exploring machine speech chain for domain adaptation," in *ICASSP*. IEEE, 2022, pp. 6757–6761.

[4] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv:1703.10135*, 2017.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*. IEEE, 2018, pp. 4779–4783.

[6] 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, 2019.

[7] 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:2006.04558*, 2020.

[8] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *ICML*. PMLR, 2021, pp. 5530–5540.

[9] Y. Liu, R. Xue, L. He, X. Tan, and S. Zhao, "Delightfultts 2: End-to-end speech synthesis with adversarial vector-quantized auto-encoders," *arXiv:2207.04646*, 2022.

[10] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification." in *Interspeech*, 2013, pp. 925–929.

[11] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv:2109.00537*, 2021.

[12] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[13] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients." in *Odyssey*, vol. 2016, 2016, pp. 283–290.

[14] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention." in *Interspeech*, 2018, pp. 681–685.

[15] W. Cai, H. Wu, D. Cai, and M. Li, "The dku replay detection system for the asvspoof 2019 challenge: On data augmentation, feature representation, classification, and fusion," *arXiv:1907.02663*, 2019.

[16] S. Chakroborty and G. Saha, "Improved text-independent speaker identification using fused mfcc & imfcc feature sets based on gaussian filter," *International Journal of Signal Processing*, vol. 5, no. 1, pp. 11–19, 2009.

[17] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[18] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "Stc antispoofing systems for the asvspoof2021 challenge," in *ASVspoof*, 2021, pp. 61–67.

[19] Y. Zhao, R. Togneri, and V. Sreeram, "Multi-task learning-based spoofing-robust automatic speaker verification system," *Circuits, Systems, and Signal Processing*, vol. 41, no. 7, pp. 4068–4089, 2022.

[20] C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal on Information Security*, vol. 2021, no. 1, pp. 1–14, 2021.

[21] A. Jain, P. Korshunov, and S. Marcel, "Improving generalization of deepfake detection by training for attribution," in *MMSP*. IEEE, 2021, pp. 1–6.

[22] R. Li, M. Zhao, Z. Li, L. Li, and Q. Hong, "Anti-spoofing speaker verification system with multi-feature integration and multi-task learning." in *Interspeech*, 2019, pp. 1048–1052.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP*. IEEE, 2021, pp. 6369–6373.

[25] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Odyssey*, 2018, pp. 74–81.

[26] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *SLT*. IEEE, 2018, pp. 1021–1028.

[27] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP*, 2022, pp. 6367–6371.