2024 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2024, LONDON, UK

IS THAT ME? USING SPEAKER IDENTITY TO DETECT FAKE SPEECH

Shilpa Chandra[⋆] *Padmanabhan Rajan*[†]

School of Computing and Electrical Engineering
Indian Institute of Technology, Mandi
[⋆]s22004@students.iitmandi.ac.in, [†]padman@iitmandi.ac.in

ABSTRACT

Of late, techniques to generate fake speech have become more and more sophisticated, resulting in challenges in their detection. This paper explores using speaker information in detecting fake speech. Speaker information derived from the linear prediction residual signal is used to supplement a state-of-the-art fake speech detector. Multi-branch convolutions followed by a transformer encoder is used to represent the residual signal compactly. Further, by novel utilization of a contrastive loss function, speaker information is captured effectively from a given utterance and an additional genuine utterance from the same speaker. On evaluation under a speaker-aware protocol, the proposed method shows promise in fake speech detection accuracy on the ASVspoof 2019 and ASVspoof 2021 datasets. Additionally, several ablation studies reveal the effectiveness of the residual signal in capturing speaker information.

***Index Terms*—** deepfake, LP residual, ASVspoof, speaker-aware

1. INTRODUCTION

Recent advances in neural text-to-speech (TTS) systems and in voice conversion have made it easy to generate malicious speech utterances. A particular category of malicious speech is *spoofed* speech, in which speech is generated to sound like a particular person. The quality of such spoofed speech has been steadily increasing, and they pose a threat to voice biometric systems such as automatic speaker verification (ASV) systems. The ASVspoof series of research challenges was one of the first efforts to compare and evaluate techniques for detecting spoofed speech [1, 2, 3, 4]. *Speech deepfakes* are malicious speech that can be used to deceive human listeners, not necessarily ASV systems. Research challenges such as ASVspoof 2021, and audio deepfake detection (ADD) [5] evaluate the performance of speech deepfake detectors [4, 5]. In this paper, the terms "spoofed speech", "speech deepfakes" and "fake speech" are used interchangeably.

This work was partially supported by the Directorate of Forensic Science Services, Government of India, under the Project - IITM/DFSS/AB/385.

Given the task of classifying a given utterance as genuine or fake, in many situations it is possible to provide additional information to the classifier. One such example is the speaker information. The question we ask in this paper is the following: Given a speech utterance x purportedly spoken by speaker s, can we use genuine utterances from s to help decide if x is genuine or fake? For example, if s is a public figure like a politician or a celebrity, it may be possible to obtain genuine utterances of s, and such a question becomes of relavance.

Many aspects of speaker identity are provided by the linear prediction (LP) residual, which is known to capture various aspects of the excitation source in the source-system model of speech. The residual has a flat spectrum, has no formant information, and incorporates the fundamental frequency F0. The LP residual has been used for speaker identification [6], and more recently for pitch estimation [7] and for modeling prosody in TTS systems [8]. In this paper, we utilize the information in the LP residual via a transformer-based architecture, and utilise a contrastive loss function to focus on the speaker identity.

Our experiments reveal that providing speaker information through the LP residual to the speech deepfake detector results in better detection accuracy. We further perform several ablation studies to infer the importance of various components of the proposed speaker-aware speech deepfake detector. We compare results to a similar study described in [9] and show improvement in detecting various types of deepfake speech.

2. RELATED WORK

Recent techniques to detect fake speech include using graph attention [10, 11, 12], conformers [13] or complex-valued spectra [14]. Most of these methods rely on finding artefacts resulting from synthetic speech generation. Tak et. al. [10] used graph attention networks which considered both spectral and temporal attention, as often the artefacts may be present in specific temporal or spectral sub-bands.

The state-of-the-art model AASIST [11] proposes the usage of concurrently operating on both spectral and temporal graphs. Moreover these graphs can be heterogeneous so that

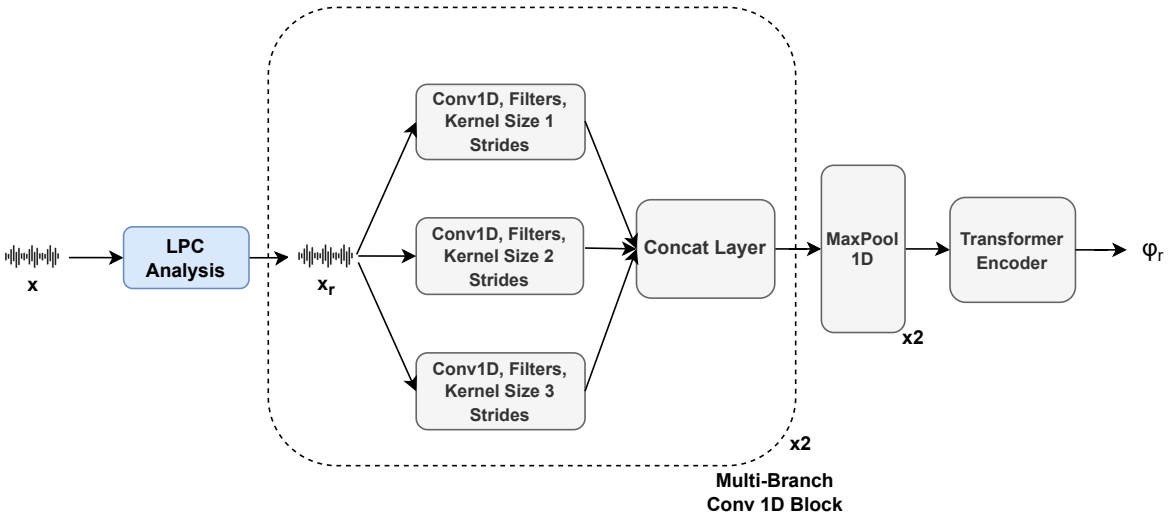


Fig. 1. LP residual network (LPRN). Blue box is non-trainable component whereas grey boxes are trainable components.

it can support different numbers of nodes, as well as different dimensionalities. These spectral and temporal graphs are then combined to form a single graph that can achieve attention spanning across both domains. Further layers are used to learn different groups of artefacts.

Another work [13] hypothesized that time-cues can be located simultaneously at multiple sub-bands. For this, the authors proposed pyramid conformers to capture both local and global information. Elastic penalty margin softmax was used to deal with unknown attacks.

Sun et. al. [15] utilized the artefacts left by different vocoders and further used this information for fake speech detection. Their model first determined the vocoder type followed by binary classification for fake speech detection. Yet another work [16] explored using artifacts present in the stereo version of the audio after mono-to-stereo conversion using a dual-branch neural architecture to process the left and right channels. Though many recent speech deepfake detectors give considerably good results, a common problem is that of generalization to completely new types of synthetic speech [17].

Similar to the present study, the paper [9] utilised speaker information to improve detection of spoofed speech. The AASIST model [11], is combined with a pre-trained speaker-embedding model ECAPA-TDNN [18] to perform speaker-aware fake speech detection.

3. USING THE LP RESIDUAL

Linear prediction (LP) is a commonly used speech analysis technique that models a sample of speech as a linear combination of the past p samples. LP analysis models the speech production system as an all-pole filter characterized by the transfer function

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{1}$$

The LP residual x_r is obtained by inverse filtering the speech x with the filter $H(z)$. Fig 2 shows the speech signal and the corresponding LP residual obtained after a 16th order LP analysis. It can be seen that several aspects of the excitation source, such as pitch period, is captured by the residual.

To capture information from the LP residual, we utilize a multi-scale convolution-based architecture with varying kernel sizes, followed by a transformer encoder layer for capturing long range dependencies. The utility of capturing long range dependencies was demonstrated by Batra et al. [7]. The LP residual from a given utterance is represented as a 512-dimensional embedding obtained after linear projection from the transformer encoder layer. We refer to this architecture as LP residual network (LPRN), which is shown in Figure 1, and detailed in Table 2.

4. USING SPEAKER INFORMATION FOR SPEECH DEEPFAKE DETECTION

We assume the availability of additional genuine utterances (henceforth called enrollment utterances) for all speakers the system works with. The protocol for evaluating the proposed detectors are described in Sec 5.1.

The AASIST model described previously represents an utterance x as a 160-dimensional embedding. To incorporate speaker information, an enrollment utterance x_e from the corresponding speaker in the form of LPRN embeddings are concatenated with AASIST embeddings. The resulting 672-dimensional embedding is speaker-dependent, in that it incorporates speaker characteristics along with spectro-temporal features that help in discriminating genuine speech from fake

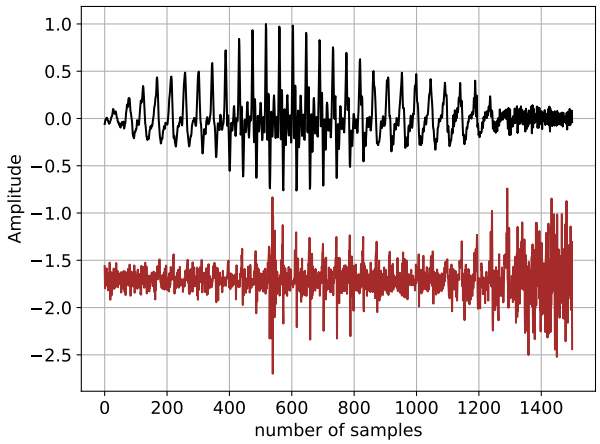


Fig. 2. A segment of speech (black) and corresponding residual (red) after LP analysis with $p = 16$. For better illustration, the amplitudes have been normalized between (-1,+1) and an offset is added to separate the plots.

speech. The embeddings are passed to classification layers, and the architecture is trained in an end-to-end fashion. This model is termed speaker-aware deepfake detector (SADD) and is shown in Figure 3 (a).

The simple concatenation of speaker characteristics, though effective, fails to explicitly associate the common speaker information in the enrollment utterance and the input utterance. To mitigate this, we use a Siamese-based LPRN network with a pair of utterances from the same speaker, one of which is the enrollment utterance. The resulting network is termed as Siamese speaker-aware deepfake detector (sSADD) and is shown in Figure 3 (b). The Siamese network uses a contrastive loss defined as

$$L_{\text{cont}}(y, x, x_e) = (y)\frac{1}{2}D + (1-y)\frac{1}{2}\max(0, m-D) \quad (2)$$

where, $D = \|\phi_{rx} - \phi_{re}\|$ is the Euclidean distance between LPRN embeddings ϕ_{rx} and ϕ_{re} of x and x_e respectively and m is the margin value which is set at 2. Here the label y is 1 for positive samples (when x is genuine) and 0 for negative samples (when x is fake). x_e by definition is always genuine.

The contrastive loss minimizes the distance between x and x_e when y is positive and maximises upto the margin m when y is negative. This encourages the network to bring together genuine embeddings in a speaker-dependent manner, at the same time pushing apart fake embeddings. Additional information from artifacts produced by fake speech are provided by concatenation with AASIST embeddings. Finally, a weighted categorical cross entropy (CCE) loss is combined with the contrastive loss, resulting in embeddings that are compact as well as discriminative. Therefore the total loss for sSADD is,

$$L_{\text{total}} = L_{\text{CCE}} + L_{\text{cont}}(y, x, x_e). \quad (3)$$

5. EXPERIMENTS

We evaluate the proposed SADD and sSADD networks using two well-known public deepfake speech corpora, namely ASVspoof 2019 logical access data, and ASVspoof 2021 logical access data [3, 4]. Since the objective is to study the performance of the detector in a speaker-dependent fashion, we follow the protocol outlined in [9], where non-relevant utterances in the dataset are discarded. This differs from the standard ASVspoof 2019 and ASVspoof 2021 protocols, which are not speaker-dependent.

5.1. Datasets and protocol used

The ASVspoof 2019 logical access data includes genuine speech, and fake speech produced as a result of voice conversion and TTS techniques (termed attacks). The dataset is partitioned into train, dev and eval sets, with mutually exclusive set of speakers in each [3]. Following [9], the train, dev and eval have 20, 10 and 48 speakers respectively, resulting in 23780 utterances in the development set and 69252 utterances in the evaluation set. There are 19 types of attacks, of which 6 are common in train and dev sets, and the remaining 13 are available only in eval, and are hence unseen.

The ASVspoof 2021 logical access data consists of the ASV2019 logical access data, under various transmission artifacts [4]. Applying the same protocol, we are left with 146829 utterances in the evaluation set. For a given utterance, the corresponding speaker information is used from the provided metadata. The enrollment utterance corresponding to a speaker is chosen randomly from the genuine utterances corresponding to that speaker.

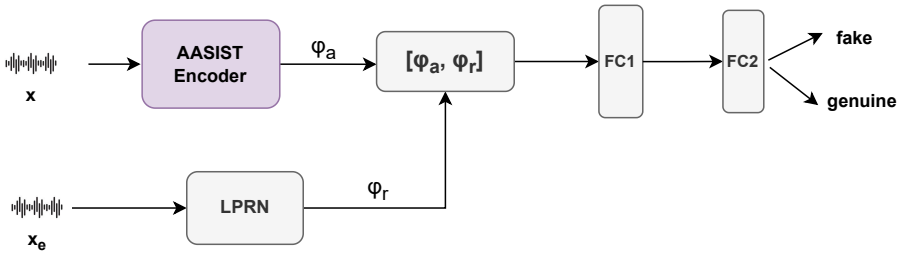
5.2. Training procedure and architecture details

The SADD network is trained with CCE loss, and the sSADD network is trained according to the loss function in Equation 3, with Adam optimizer and cosine annealing learning rate decay. Due to hardware constraints, batch sizes of 32 and 16 were used for SADD and sSADD respectively, while AASIST was kept fixed during training. We utilized the Nvidia 2080x Ti GPU (12GB RAM).

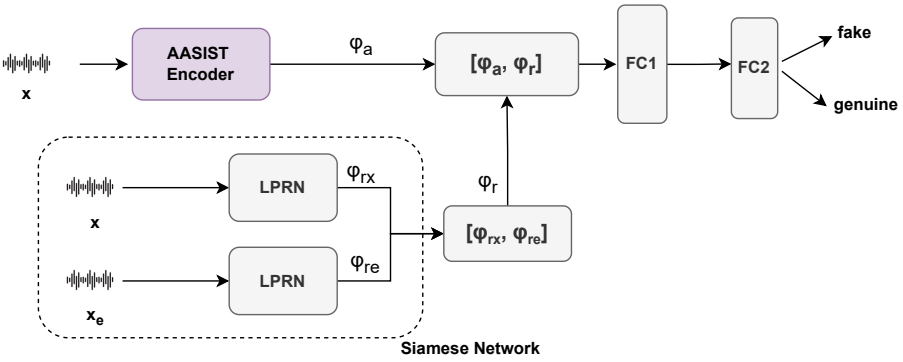
The LPRN network uses the Multi-Branch Conv1D Block twice in a serial manner. Maxpool1D is used twice, and transformer encoder, which has two layers and a 512-dimensional feed forward network, is used as the last step. The Table 2 contains details of all modules used in LPRN, SADD and sSADD networks.

5.3. Results

We evaluate SADD and sSADD models using the protocol described above. For ASVspoof 2019 data, results are presented in terms of equal-error rate (EER) and pooled tandem detection cost function (t-DCF), and for ASVspoof 2021 data, results are presented in terms of EER as described in [3, 4]. For ASVspoof 2019 we compare results with [9] which, to the



(a) Speaker-aware deepfake detector(SADD).



(b) Siamese speaker-aware deepfake detector (sSADD).

Fig. 3. Speaker-aware networks. Due to hardware limitations, the purple box is not trained whereas grey boxes are trained. x is the input utterance, x_e is the enrollment utterance, ϕ_a represents embedding from AASIST, ϕ_r represents LP residual embedding, [a,b] represents concatenation of a and b.

Table 1. Breakdown EER (%) performance of all 13 attacks that exist in the ASVspoof 2019 LA evaluation set, pooled min t-DCF (P1), and pooled EER (%, P2).

Method	**A07**	**A08**	**A09**	**A10**	**A11**	**A12**	**A13**	**A14**	**A15**	**A16**	**A17**	**A18**	**A19**	**P1**	**P2%**
sSADD	0.34	0.26	0.00	0.59	0.09	0.40	0.05	0.07	0.36	0.46	1.04	2.17	0.42	0.017	0.63
Liu et al. [9]	1.18	0.07	0.00	1.38	0.41	0.98	0.22	0.28	0.98	0.65	1.28	2.70	0.34	0.038	1.13

best of our knowledge is the only other work to use speaker-aware fake speech detection.

Table 4 shows that the proposed methods give excellent performance, with sSADD perfoming a notch better. For ASVspoof 2019, an overall relative improvement of 52.35% and 44.24% in EER is obtained when compared to non speaker-aware [11] and speaker-aware [9] settings, evaluated according to the protocol in [9]. For ASVspoof 2021, we compare results with non speaker-aware model [11], obtaining a modest relative improvement of 2.5% as seen in Table 4. The sSADD performs slightly better than SADD model.

The results obtained in Table 4 are without any data augmentation techniques. As can be seen from the table, additional speaker-specific information is available to sSADD over SADD. These results also demonstrate the effectiveness of using speaker information available in the LP residual. The embedding plot in Figure 4 illustrates how the sSADD network can distinguish spoof samples from genuine samples in a distinctive manner. Table 1 also displays a performance breakdown by attack for ASVspoof 2019. It is evident that nearly all evaluation attacks in the evaluation set spanning from A07 to A19 have significantly decreased. Furthermore, the infamous A18 attack has decreased in comparison to the approach in [9]. Notably, we are limited in our comparison with other approaches for ASVspoof 2019 and 2021 datasets since we are constrained by the speaker-aware protocol [9].

5.4. Ablation studies

The previous sections show the effectiveness of using speaker information from enrollment utterances in detecting fake speech. We conduct a number of ablation studies to evaluate the contribution of various components in the proposed SADD and sSADD systems. The results on ASVspoof 2019

Component	Description
Multi-Branch Conv1D Block A (LPRN)	Filters: 4, Strides: 2, Kernel Size1: 7, Kernel Size2: 11, Kernel Size3: 17
Multi-Brach Conv1D Block B (LPRN)	Filters: 4, Strides: 2, Kernel Size1: 5, Kernel Size2: 7, Kernel Size3: 9
MaxPool1D (LPRN)	Kernel Size: 2
Tranformer Encoder (LPRN)	Encoder Layers: 2, Attention heads: 2, Fully Connected dim: 512
Fully Connected Layer 1 (SADD& sSADD)	Fully Connected dim: 256
Fully Connected Layer 2 (SADD& sSADD)	Fully Connected dim: 2

Table 2. Detailed description for components of LPRN, SADD and sSADD networks.

(in terms of EER) are presented in Table 3.

5.4.1. Using information of another speaker

The performance of the proposed SADD and sSADD models are evaluated when mismatched speaker information (in other words, using enrollment utterance from a different speaker than the one present in x) is used. The results reveal, rather surprisingly, that there is only a slight drop in performance when compared to using the correct speaker information. Further studies using a larger pool of speakers is needed to interpret this result.

5.4.2. Using alternate embeddings for speaker information

We also evaluate the performance of SADD and sSADD using pre-trained ECAPA-TDNN embeddings in place of LPRN embeddings. ECAPA-TDNN embeddings have shown excellent performance in tasks such as speaker identification. The results in Table 3 show that ECAPA-TDNN embeddings perform poorer than LPRN features in providing speaker information especially in the sSADD network. This might be the case because only genuine speech utterances are used to train the ECAPA-TDNN network. However, the LPRN network is trained on both real and fake samples, and its embeddings are obtained from a low-level representation of the excitation source, which results in a more accurate representation of the speaker's features.

Table 3. Performance in terms of EER for the ablation studies.

Method	SADD-EER	sSADD-EER
Mismatched speaker	0.70	0.64
with ECAPA-TDNN	1.13 [9]	3.03
without AASIST	> 30	> 30

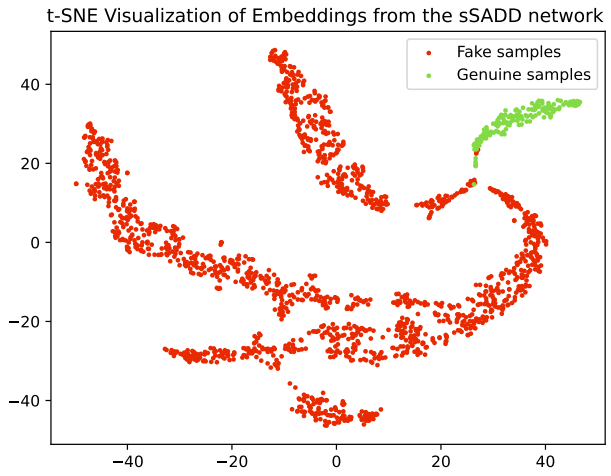


Fig. 4. t-SNE visualization of embeddings generated by sSADD network of a single speaker on the ASVspoof 2019 dataset.

5.4.3. Not using AASIST features

Finally, we also evaluate the importance of AASIST embeddings in the overall pipeline. We only pass the LPRN embeddings to the SADD and sSADD models and evaluate the performance. This resulted in EERs in the range of 35%. This implies that AASIST embeddings are crucial for the successful detection of fake speech in both SADD and sSADD models. This is not surprising, since LPRN embeddings are not designed to capture artifacts present in fake speech. The LP residual discards the system information, which has important contributions in speech representation.

Table 4. Results in terms of EER and t-DCF on the datasets described in Section 5.1.

Methods	ASVspoof 2019		ASVspoof 2021
Non Speaker-Aware Methods	EER	t-DCF	EER
AASIST	1.32	0.040	7.79
Speaker-Aware Methods	EER	t-DCF	EER
Liu et al. [9]	1.13	0.038	-
SADD (proposed)	0.68	0.018	7.64
sSADD (proposed)	0.63	0.017	7.59

6. CONCLUSION

In this paper, we incorporated speaker information from the linear prediction residual to aid in the detection of fake speech. Combining speaker information with features from the AASIST detector resulted in improved detection performance. The improved performance was obtained by utilizing the common speaker information present in an evaluation utterance and a genuine utterance from the same speaker. This

study reveals that although AASIST provides the bulk of the performance, there is merit in using simple pre-processing techniques (here, linear prediction of speech) to feed useful information to aid powerful learning models. Future investigations will include a larger pool of speakers to evaluate the effectiveness of using speaker information to detect fake speech.

7. REFERENCES

[1] Z. Wu at al., "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH 2015*, 2015, pp. 2037–2041.

[2] Tomi Kinnunen, Md. Sahidullah, and Héctor Delgado et al., "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Proc. INTERSPEECH 2017*, 2017, pp. 2–6.

[3] Massimiliano Todisco, Xin Wang, Ville Vestman, and Md. Sahidullah et al., "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. INTERSPEECH 2019*, 2019, pp. 1008–1012.

[4] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[5] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al., "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[6] SR Mahadeva Prasanna, Cheedella S Gupta, and B Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

[7] Mudit D Batra, MK Jayesh, and CS Ramalingam, "Robust pitch estimation using multi-branch cnn-lstm and 1-norm lp residual.," in *INTERSPEECH*, 2022, pp. 3573–3577.

[8] Zhao-Ci Liu, Zhen-Hua Ling, Ya-Jun Hu, Jia Pan, Jin-Wei Wang, and Yun-Di Wu, "Speech Synthesis with Self-Supervisedly Learnt Prosodic Representations," in *Proc. INTERSPEECH 2023*, 2023, pp. 7–11.

[9] Xuechen Liu, Md Sahidullah, Kong Aik Lee, and Tomi Kinnunen, "Speaker-Aware Anti-spoofing," in *Proc. INTERSPEECH 2023*, 2023, pp. 2498–2502.

[10] Hemlata Tak, Jee-weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas Evans, "Graph attention networks for anti-spoofing," *arXiv preprint arXiv:2104.03654*, 2021.

[11] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

[12] Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *arXiv preprint arXiv:2107.12710*, 2021.

[13] Jingran Gong and Ning Chen, "Synthetic Voice Spoofing Detection based on Feature Pyramid Conformer," in *Proc. INTERSPEECH 2023*, 2023, pp. 2803–2807.

[14] Nicolas M. Müller, Philip Sperl, and Konstantin Böttinger, "Complex-valued neural networks for voice anti-spoofing," in *Proc. INTERSPEECH 2023*, 2023, pp. 3814–3818.

[15] Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu, "Ai-synthesized voice detection using neural vocoder artifacts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 904–912.

[16] Rui Liu, Jinhua Zhang, Guanglai Gao, and Haizhou Li, "Betray Oneself: A Novel Audio DeepFake Detection Model via Mono-to-Stereo Conversion," in *Proc. INTERSPEECH 2023*, 2023, pp. 3999–4003.

[17] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger, "Does audio deepfake detection generalize?," *arXiv preprint arXiv:2203.16263*, 2022.

[18] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. INTERSPEECH 2020*, 2020, pp. 3830–3834.