PHASE-AWARE SPOOF SPEECH DETECTION BASED ON RES2NET WITH PHASE NETWORK

Juntae Kim\textsuperscript{1}, Sung Min Ban\textsuperscript{1}

\textsuperscript{1}SK Telecom, Seoul, Republic of Korea

\section*{ABSTRACT}

For automatic speaker verification systems, spoof speech detection (SSD) is an essential countermeasure. Although SSD with magnitude features in the frequency domain has shown promising results, phase information can also be useful in capturing the artefacts of certain spoofing attacks. Thus, both magnitude and phase features must be considered to ensure the ability to generalize diverse types of spoofing attacks. In this study, we discovered that the randomness difference between magnitude and phase features is large, which can interrupt the feature-level fusion via backend neural network. In this regard, we propose a phase network to reduce that difference, which makes the Res2Net-based feature-level fusion feasible. To validate our SSD system for practical environment, both known- and unknown-type SSD scenarios are considered. As a result, our SSD system delivers competitive results compared to other state-of-the-art SSD systems in all scenarios.

\textbf{Index Terms}— anti-spoofing, phase

\section{1. INTRODUCTION}

The automatic speaker verification (ASV) system \cite{1} has recently become an attractive research topic because the ASV system has been widely applied in commercial applications. However, ASV technologies have long been known to be vulnerable to spoofing attacks. Thus, ASV countermeasure technologies, such as spoof speech detection (SSD), which classifies between bona fide (genuine) audios and spoof audios, must be equipped.

Spoofing attacks are mainly categorized into two types: physical access (PA) and logical access (LA). The PA considers the replay attack \cite{2, 3, 4}, whereas spoofing attacks using text-to-speech synthesis \cite{5, 6, 7, 8} and voice conversion \cite{9, 10, 11} technologies are considered in LA. Due to the diverse types of spoofing attacks, the generalized ability of various unseen spoofing attacks is the most crucial attribute for building the SSD system.

We discovered that most recent SSD studies are based on ASVspoof2019 \cite{12} in which numerous studies have investigated the robustness of various acoustic features to PA and LA using the state-of-the-art backend neural networks, such as ResNet \cite{13, 14, 15, 16}, Res2Net \cite{17, 18}, and graph attention networks \cite{19, 20}. The log-power (or magnitude) discrete Fourier transform (DFT), constant-Q transform (CQT), and linear frequency cepstral coefficients (LFCC), which are the magnitude features in the frequency domain \cite{16, 17, 21, 22}, are representative acoustic features for SSD. Among them, log-power CQT-based approaches have shown outstanding performance both in PA and LA because CQT effectively represents low and high frequency bands with multiple frame windows with difference lengths, implying that CQT has fluent spectral information to capture the spoofing cues \cite{17}.

Although approaches based on magnitude features have shown promising results, there is a report that the use of phase information can be important in detecting some types of voice conversion-based spoofing attacks: A17 attack in ASVspoof2019, which is difficult to detect using magnitude features \cite{23}. Considering the generalization ability to deal with various spoofing attacks, using both magnitude and phase features effectively is a crucial research topic.

In this study, the feature-level fusion of magnitude and phase features over Res2Net was achieved to boost the complementary effect of those features by adopting an additional phase network that reduces the randomness difference between magnitude and phase features before feeding them into Res2Net. Considering the practical applications, the proposed approach was validated in both known- and unknown-type SSD scenarios, where the type includes PA and LA because the specific type of spoofing attacks cannot be known in practical environment. Experiments exhibit the following: (i) outperforming other feature- and score-level strategies in a great margin for LA (e.g., A17 attack), while showing competitive results for PA in all scenarios; (ii) showing the competitive results compared to other state-of-the-art SSD systems in all scenarios.

\section{2. PROPOSED METHOD}

\subsection{2.1. Motivation}

Some studies \cite{13, 16, 19, 24, 25} have used phase information for SSD. \cite{19} used the raw waveform to deploy phase information with a backend neural network named AASIST, which showed extraordinary performance in LA but not in PA. \cite{13} proposed CQT with modified magnitude-phase spectrum (MMPS) to hand-craft the integration of the CQT-magnitude and -phase spectra. Although the CQT-MMPS outperformed the CQT-power spectrum in PA and LA, there is no guarantee that it is an optimal method for combining the CQT-magnitude and -phase spectra for the synergy effect. Additionally, no studies into A17 have been conducted to demonstrate the effectiveness of phase information.

To boost the synergy effect between magnitude and phase spectra, one option is to train a neural network with raw magnitude and phase spectra, with the belief that the neural network will effectively use the SSD-related information from magnitude and phase spectra in a data-driven way. \cite{16} concatenated the raw magnitude and phase spectra across the channel dimension, and then directly fed into the input convolution layer of ResNet to use the complementary phase information. Although individual magnitude and phase spectra were shown that they have respective discriminative power between bona fide and spoofed speech in PA, aggregating them did not show the synergy effect, even degrading the detection performance compared with the model using the magnitude spectrum only. Furthermore, no study on the effectiveness of phase information for LA has been done. Owing to the failure of feature-level fusion between magni-
Neural Network
Magnitude
Backend

(c) Score weighted sum
Backend
Neural Network
Magnitude
Features
Phase Features

(b) Concatenation
Bonafide / Spoof
Weighted Sum

Backend
Neural Network
Magnitude
Features
Phase Features

(c) Score weighted sum
Backend
Neural Network
Magnitude
Features
Phase Features

(d) Phase network and concatenation
Bonafide / Spoof
Batch Normalization
ReLU
Phase network

Fig. 1. CQT-magnitude spectrum in dB scale (up) and the entropy analysis (down) for the CQT-magnitude spectrum (red), CQT-phase spectrum (green), random noise image (blue), and phase network output features (yellow).

We hypothesize that the failure of neural network-based feature-level fusion of magnitude and phase features is caused by the different feature complexity levels of magnitude and phase spectra. Because the phase spectrum has an inevitable wrapping problem [26], resulting in a random pattern, analyzing some patterns from the phase spectrum is more difficult than analyzing patterns from the magnitude spectrum, which has well-structured patterns, such as formant and harmonicity.

To compare the randomness of the magnitude and phase spectra, the CQT-magnitude and -phase spectra were subjected to an entropy analysis, as shown in Fig. 1. After global min-max normalization, the entropy of each time frame is calculated for the entropy analysis. We also analyze the entropy of random noise image to compare its randomness with that of CQT-phase spectrum. Note that each pixel of random noise image is independently sampled from the uniform random variable $X \sim U (0, 1)$. Fig. 1 shows that the entropy levels of the CQT-phase spectrum and random noise image are similar, whereas the entropy level of the CQT-magnitude spectrum is lower than the formers, particularly in the voice activity region (1.5-3.0 s), implying that the CQT-phase spectrum has higher randomness than the CQT-magnitude spectrum. Note that the phase wrapping problem exists regardless of the voice activity region; thus, the entropy level of CQT-phase spectrum is consistent for all time frames.

The high randomness of the phase spectrum can be a possible reason for the failure of concatenation-based feature-level fusion because the magnitude spectrum can overwhelm the training of the neural network with magnitude and phase spectra, causing the additional model parameters for the phase spectrum to be meaningless, which suffers the model’s generalization ability, which is important for SSD [17], and thus degrading the detection performance. This motivated us to propose the additional processing for the phase spectrum only, to reduce its randomness and balance the entropy level of magnitude and phase spectra before feeding them to the backend neural network model. Considering that the RGB image-based ap-

2.2. Frameworks for the phase-awareness
To validate the effectiveness of using phase information, we investigate four frameworks, as shown in Fig. 2(a)–2(d), where ‘⊕’ implies the concatenation across the channel dimension. The framework (a) uses magnitude features alone, (b) directly concatenates magnitude and phase features across the channel dimension, (c) uses the magnitude and phase features to train each individual backend neural network, after which the final score is achieved by applying a weighted sum method for the scores from backend neural networks, where the score weights are 1.0 and 0.02 for magnitude- and phase-based backend neural network, respectively (found from the development set), (d) firstly processes the phase features by a phase network in the framework, after which the processed phase and magnitude features are concatenated across the channel dimension.

Considering the objective of the phase network is to extract features and reduce the entropy level of the original phase features, the phase network mainly consists of convolution layers. Note that the convolution layer has a feature extraction ability [31]. The details
of the phase network are described in Fig. 2(d). For the convolution block (Conv2d), the first two numbers in parentheses indicate the channel size and kernel size, respectively. The corresponding dimension for the first and second number in the kernel size is time and feature dimension, respectively. The aforementioned configuration of the phase network was found from the development set.

Res2Net was used as the backend neural network. Res2Net50 with squeeze-and-excitation block introduced in [17] was used as Res2Net with the same configuration. We refer to Res2Net with frameworks (a), (b), (c) and (d) in Fig. 2, as Res2Net, Res2Net-FF (feature-level fusion), Res2Net-SF (score-level fusion), and Res2Net-FF+PN (feature-level fusion with phase network) respectively. The input channel dimension for Res2Net and Res2Net-SF is 1, that for Res2Net-FF and Res2Net-FF+PN is 2. All Res2Net-based frameworks have approximately 0.84M model parameters excepting for the Res2Net-SF which has double parameters that of Res2Net. As the phase network is a shallow convolution neural network, the parameter number difference between Res2Net and Res2Net-FF+PN is approximately 200, which is a minor number considering the total number of model parameters.

### 3. EXPERIMENTS AND RESULTS

#### 3.1. Experimental setup

**Dataset:** The ASVspoof 2019 challenge dataset, which contains two subsets: PA and LA, was used for all experiments. Each subset includes training, development, and evaluation partitions. The training and development partitions were used for model training and selection, respectively, while the evaluation partition was used for model evaluation. More details about the dataset used can be found in [12].

**Evaluation metrics:** The trained models are evaluated using the tandem detection cost function (t-DCF) [32] and equal error rate (EER) [32]. The log-probability of the bona fide class was used as the score for t-DCF and EER computation.

**Feature specification:** The CQT was extracted using a 16 ms hop length, a Hanning window, and nine octaves with 12 bins per octave, resulting in a feature dimension of 108 for CQT.

**Feature engineering:** As speech utterances differ in length, some feature engineering is necessary. In this study, the unified feature map method in [15] was used. Therefore, we first extended all utterances to a multiple of 400 frames. After that, the extended utterance was divided into length 400-frame segments with a 200-frame overlap. The backend neural network outputs from multiple segments were simply averaged for each utterance during evaluation. More details about the dataset used can be found in [12].

**Training and evaluation scenario:** We discovered that most SSD studies follow the known-type SSD scenario, which means that if a model is trained on PA, it is evaluated only in PA. However, the unknown-type SSD scenario, in which the model is trained on both PA and LA to detect spoof speech regardless of spoof type, is the most practical scenario. This study includes both known- and unknown-type SSD scenarios to validate the proposed method in a real-world setting.

**Training strategy:** The training strategy is similar to [17]. The binary cross entropy loss was used to train all models. Adam is adopted as the optimizer with \( \beta_1 = 0.9, \beta_2 = 0.98 \), and weight decay of 10\(^{-5}\). All models are trained for 30 epochs, and the model with the lowest EER on the development set is chosen for evaluation. Furthermore, we discovered that the SSD performance using ASVspoof 2019 has high variation according to the initialization weights [21]. To strictly validate the proposed method, all experiments were repeated three times with different random seeds.

#### 3.2. Experimental results and discussion

Table 1 compares four types of Res2Net frameworks for the known-type SSD scenario. In LA evaluation, we discovered that Res2Net-FF+PN outperformed all other methods by a great margin in both t-DCF and EER. Conversely, Res2Net-FF performed worse than the Res2Net. The Res2Net-SF marginally outperformed Res2Net. In PA evaluation, both Res2Net and Res2Net-SF showed the best performance and no significant performance improvement between Res2Net and Res2Net-FF+PN was observed, implying that phase information is less important for PA than LA. However, Res2Net-FF+PN still outperformed all other methods by a great margin in both PA and LA.

![Table 1](image.png)

**Table 1.** ASVspoof2019 PA and LA results in a known-type SSD scenario for four types of Res2Net frameworks. Performance reported in “average (best)” after three repeated experiments with different random seeds. The numbers in bold indicate the best-averaged result.

| Model          | Feature | PA      | LA      |
|----------------|---------|---------|---------|
|                |         | t-DCF EER(%) | t-DCF EER(%) |
| Res2Net        | CQT     | 0.007(0.005) | 0.029(0.027) |
| Res2Net-FF     | CQT+phase | 0.007(0.006) | 0.013(0.015) |
| Res2Net-SF     | CQT+phase | 0.020(0.024) | 0.050(0.052) |
| Res2Net-FF+PN  | CQT+phase | 0.030(0.036) | 0.250(0.254) |

Table 2 compares four types of Res2Net frameworks for the known-type SSD scenario in terms of t-DCF and pooled EER (%) for state-of-the-art systems and our proposed best system. The numbers in bold indicate the best result.

![Table 2](image.png)

**Table 2.** ASVspoof2019 PA and LA results on their evaluation partitions in a known-type SSD scenario in terms of t-DCF and pooled EER (%) with the lowest EER on the development set is chosen for evaluation. Furthermore, we discovered that the SSD performance using ASVspoof 2019 has high variation according to the initialization weights [21]. To strictly validate the proposed method, all experiments were repeated three times with different random seeds.
Table 3. Breakdown EER (%) performance of all 13 attacks from ASVspoof2019 LA-Evaluation, pooled min t-DCF(P1) and pooled EER(%), P2). The reported results are from the best performed models shown in Table 1. The numbers in bold indicate the best result.

| Model          | Feature   | A07 | A08 | A09 | A10 | A11 | A12 | A13 | A14 | A15 | A16 | A17 | A18 | A19 | P1  | P2  |
|----------------|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Res2Net        | CQT       | 0.16| 1.73| 0.02| 0.51| 0.24| 0.10| 0.12| 0.19| 0.47| 0.55| 2.99| 2.62| 0.88| 0.038| 1.31|
| Res2Net-FF     | CQT+phase | 0.19| 0.90| 0.02| 0.41| 0.22| 0.11| 0.12| 0.16| 0.20| 0.22| 3.82| 2.60| 1.16| 0.046| 1.47|
| Res2Net-SF     | CQT+phase | 0.16| 1.67| 0.02| 0.49| 0.24| 0.10| 0.12| 0.19| 0.47| 0.53| 2.73| 2.34| 0.86| 0.037| 1.29|
| Res2Net-FF+PN  | CQT+phase | 0.22| 2.67| 0.02| 0.51| 0.33| 0.18| 0.06| 0.22| 0.22| 0.34| 1.75| 1.77| 0.33| 0.027| 0.94|

Table 4. ASVspoof2019 PA and LA results in an unknown-type SSD scenario. Performance reported in “average (best)” after three repeated experiments. The numbers in bold indicate the best-averaged result.

| Model          | Feature   | Development | PA         | Evaluation | Development | LA         | Evaluation |
|----------------|-----------|-------------|------------|------------|-------------|------------|------------|
|                |           | t-DCF      | EER(%)     | t-DCF      | EER(%)      | t-DCF      | EER(%)     |
| Res2Net        | CQT       | 0.009(0.007)| 0.35(0.28) | 0.016(0.012)| 0.61(0.45) | 0.015(0.013)| 0.46(0.40) | 0.111(0.097)| 3.84(3.66) |
| Res2Net        | CQT-MMPS  | 0.014(0.008)| 0.53(0.31) | 0.018(0.010)| 0.66(0.40) | 0.036(0.034)| 1.16(1.09) | 0.147(0.128)| 6.27(5.51) |
| Res2Net-FF     | CQT+phase | 0.011(0.009)| 0.45(0.30) | 0.019(0.016)| 0.69(0.54) | 0.009(0.008)| 0.31(0.27) | 0.112(0.105)| 4.54(4.09) |
| Res2Net-SF     | CQT+phase | 0.012(0.009)| 0.43(0.29) | 0.022(0.017)| 0.72(0.65) | 0.008(0.007)| 0.34(0.30) | 0.109(0.103)| 4.23(4.02) |
| Res2Net-FF+PN  | CQT+phase | 0.009(0.008)| 0.39(0.31) | 0.016(0.013)| 0.58(0.49) | 0.012(0.007)| 0.38(0.23) | 0.093(0.092)| 3.33(3.23) |

In this work, the feature-level fusion of magnitude and phase features via Res2Net was achieved for generalized SSD by adopting a phase network, which was used for individual phase feature processing to reduce intrinsic high randomness of phase features. Although the proposed method performed well for PA and LA in known-type scenarios, particularly for A17; however, there is a performance degradation in unknown-type scenarios for LA compared to known-type scenarios. Considering the practical SSD systems, the performance in unknown-type scenarios is important; thus, performance reconstruction for LA in unknown-type scenarios should be considered in our future study.
challenge: Assessing the limits of replay spoofing attack detection," 2017.

[3] C. H. You and J. Yang, “Device feature extraction based on parallel neural network training for replay spoofing detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2308–2318, 2020.

[4] J. Yang, L. Xu, and B. Ren, “Constant-q deep coefficients for playback attack detection,” IEEE Transactions on Information and Systems, vol. 103, no. 2, pp. 464–468, 2020.

[5] H. Zen, M. J. Gales, Y. Nankaku, and K. Tokuda, “Product of experts for statistical parametric speech synthesis,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 3, pp. 794–805, 2011.

[6] X. Wang, S. Takaki, J. Yamagishi, S. King, and K. Tokuda, “A vector quantized variational autoencoder (vq-vae) autoregressive neural f_0 model for statistical parametric speech synthesis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 157–170, 2019.

[7] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, “Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation,” arXiv preprint arXiv:2106.07889, 2021.

[8] J. Kim, H. Choi, J. Park, S. Kim, J. Kim, and M. Hahn, “Korean singing voice synthesis system based on an lstm recurrent neural network,” in Proc. Interspeech, 2018, pp. 1551–1555.

[9] X. Tian, S. W. Lee, Z. Wu, E. S. Chng, and H. Li, “An exemplar-based approach to frequency warping for voice conversion,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 10, pp. 1863–1876, 2017.

[10] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, “Sequence-to-sequence acoustic modeling for voice conversion,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 3, pp. 631–644, 2019.

[11] B. Sisman, M. Zhang, and H. Li, “Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, pp. 1085–1097, 2019.

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

[13] J. Yang, H. Wang, R. K. Das, and Y. Qian, “Modified magnitude-phase spectrum information for spoofing detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1065–1078, 2021.

[14] Y. Yang, H. Wang, H. Dinkel, Z. Chen, S. Wang, Y. Qian, and K. Yu, “The sjtu robust anti-spoofing system for the asvspoon 2019 challenge,” in Interspeech, 2019, pp. 1038–1042.

[15] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, “Assert: Anti-spoofing with squeeze-excitation and residual networks,” arXiv preprint arXiv:1904.01120, 2019.

[16] J.-w. Jung, H.-j. Shim, H.-S. Heo, and H.-J. Yu, “Replay attack detection with complementary high-resolution information using end-to-end dnn for the asvspoon 2019 challenge,” arXiv preprint arXiv:1904.10134, 2019.

[17] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, “Replay and synthetic speech detection with res2net architecture,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6354–6358.

[18] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, “Channel-wise gated res2net: Towards robust detection of synthetic speech attacks,” arXiv preprint arXiv:2107.08803, 2021.

[19] 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,” arXiv preprint arXiv:2110.01200, 2021.

[20] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, “End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection,” arXiv preprint arXiv:2107.12710, 2021.

[21] X. Wang and J. Yamagishi, “A comparative study on recent neural spoofing countermeasures for synthetic speech detection,” arXiv preprint arXiv:2103.11326, 2021.

[22] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, “Spoofing attack detection using the non-linear fusion of sub-band classifiers,” arXiv preprint arXiv:2005.10393, 2020.

[23] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, “End-to-end anti-spoofing with rawnet2,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6369–6373.

[24] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, “Spoof detection using source, instantaneous frequency and cepstral features,” in Interspeech, 2017, pp. 22–26.

[25] R. K. Das and H. Li, “Instantaneous phase and excitation source features for detection of replay attacks,” in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018, pp. 1030–1037.

[26] K. Vijayan and K. S. R. Murty, “Analysis of phase spectrum of speech signals using allpass modeling,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 12, pp. 2371–2383, 2015.

[27] M. Schwarz, H. Schulz, and S. Behnke, “Rgd-b object recognition and pose estimation based on pre-trained convolutional neural network features,” in 2013 IEEE international conference on robotics and automation (ICRA). IEEE, 2015, pp. 1329–1335.

[28] T. Stiebel, S. Koppers, P. Seltsam, and D. Merhof, “Reconstructing spectral images from rgb-images using a convolutional neural network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 948–953.

[29] J. Tribolo, “A new phase unwrapping algorithm,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 25, no. 2, pp. 170–177, 1977.

[30] K. Steiglitz and B. Dickinson, “Phase unwrapping by factorization,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 30, no. 6, pp. 984–991, 1982.

[31] Y. Bengio, I. Goodfellow, and A. Courville, Deep learning. MIT press Cambridge, MA, USA, 2017, vol. 1.

[32] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, “Asvspoon 2019: Future horizons in spoofed and fake audio detection,” arXiv preprint arXiv:1904.05441, 2019.

[33] Z. Wang and J. H. Hansen, “Audio anti-spoofing using a simple attention module and joint optimization based on additive angular margin loss and meta-learning,” in Proc. Interspeech, 2022.

[34] A. Gomez-Alanis, J. A. Gonzalez-Lopez, and A. M. Peinado, “A kernel density estimation based loss function and its application to asv-spoofing detection,” in Proc. Interspeech, 2019.

[35] M. Alzantot, Z. Wang, and M. B. Srivastava, “Deep residual neural networks for audio spoofing detection,” arXiv preprint arXiv:1907.00501, 2019.

[36] Y. Zhang, J. Huang, J.-s. Heo, and P. Zhang, “The effect of silence and dual-band fusion in anti-spoofing system,” 2021.