Robust Audio Deepfake Detection using Ensemble Confidence Calibration

Chin Yuen Kwok[†‡*], *Duc-Tuan Truong*[‡*], *and Jia Qi Yip*[‡],

[†]Digital Trust Centre, Nanyang Technological University, Singapore
[‡]College of Computing and Data Science, Nanyang Technological University, Singapore
kwok0062@e.ntu.edu.sg

***Abstract*—Model ensembles using linear interpolation are commonly employed to improve classification performance, with higher weights assigned to better-performing models in the ensemble. However, prior methods use fixed weights across all test samples, which is suboptimal as different models may perform better in different subsets of the samples, especially in out-of-domain (OOD) scenarios. This is a key challenge in Audio Deepfake Detection (ADD) due to variations between training and testing domains. To address this, we propose using EOW-Softmax, a method for modeling open-world uncertainties, to calibrate the magnitudes of OOD classification scores at the sample level. This dynamic adjustment improves ensemble predictions on OOD samples. When tested on the ASVspoof 2021 dataset, our calibrated ensemble reduced the equal error rate (EER) from 2.66% to 2.03%.**

***Index Terms*—Audio deepfake detection, confidence calibration, model ensemble, score-level fusion, out-of-domain classification**

I. Introduction

Audio deepfake detection (ADD) is the task of distinguishing genuine utterances from synthesized ones. Recently, synthesized audio has become harder to distinguish due to the rapid advancement of audio synthesis technologies, and ADD models need to quickly learn from the audio patterns of newly developed audio synthesizers to accurately detect fake audios. However, new audio synthesizers are being trained everyday, and it is difficult for an ADD model to have learned all the audio patterns of every existing and future audio synthesizer. As such, detecting fake audio generated from synthesizers not used in ADD model training, known as the out-of-domain (OOD) scenario, is essential. However, current ADD models still struggle to perform in the OOD scenario [1].

To improve the robustness of ADD models in OOD, previous work forms model ensembles and combines the classification score of multiple ADD models [2], [3]. In general, two types of score-level fusion methods have been studied to obtain optimal score combinations. The first method is based on the likelihood ratio test, but this does not apply to the OOD scenario as the test data distribution is not known [4]. This paper focuses on the second method, which can be used in the OOD scenario. They are linear fusion algorithms [5]–[7] like average, weighted average, maximum, minimum, and median score-level fusions. These methods, which combine multiple scores linearly, have been shown to consistently improve the performance of ADD model ensembles.

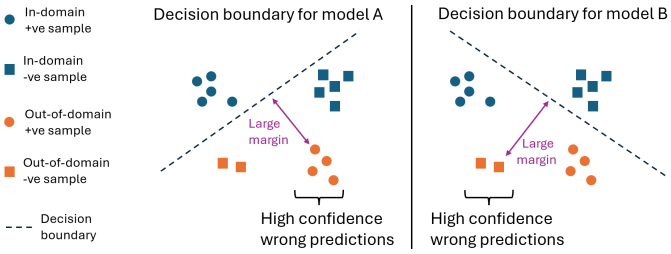

Fig. 1. Overview of the inconsistent performance across different subsets of the out-of-domain (OOD) samples. Given two models A and B in a model ensemble, a different decision boundary is learnt for each model. For model A on the left, it performs better in identifying the OOD negative samples compared to the OOD positive samples, and vice versa for model B on the right. This shows that a model's performance can vary across different subsets of OOD samples. Furthermore, the large margin from the decision boundary shows that the models are often overconfident of the wrong predictions for OOD samples, and may produce a larger magnitude for the wrong classification scores and dominate the interpolated ensemble output. Therefore, dynamic adjustments should be made to the magnitudes of the model outputs to improve the ensemble output combination.

The effectiveness of the methods lies in their ability to adjust the scores' magnitudes during the output combination. By scaling the classification scores of more likely correct predictions to a larger magnitude, their influence in the final ensemble output increases. However, these methods still struggle to adjust the score magnitudes of the OOD predictions. This is because the optimal scaling found during validation does not adapt to the OOD test set.

To improve the magnitude adjustment in OOD, we proposed using the Energy-based Open-World Softmax (EOW Softmax) [8], a method for modeling open-world uncertainties, to perform OOD confidence calibration on ADD model ensembles. Confidence calibration is the process of adjusting the predicted probabilities of a model to better reflect the true likelihood of its predictions being correct [9], which is critical for reliable decision-making in applications like medical diagnosis or autonomous systems. By adjusting the probabilities, the logit scores, which are the outputs of ADD models before Softmax is applied, also have their magnitudes adjusted, and we show that these adjusted magnitudes can improve the ensemble result.

[*]These authors contributed equally to this work

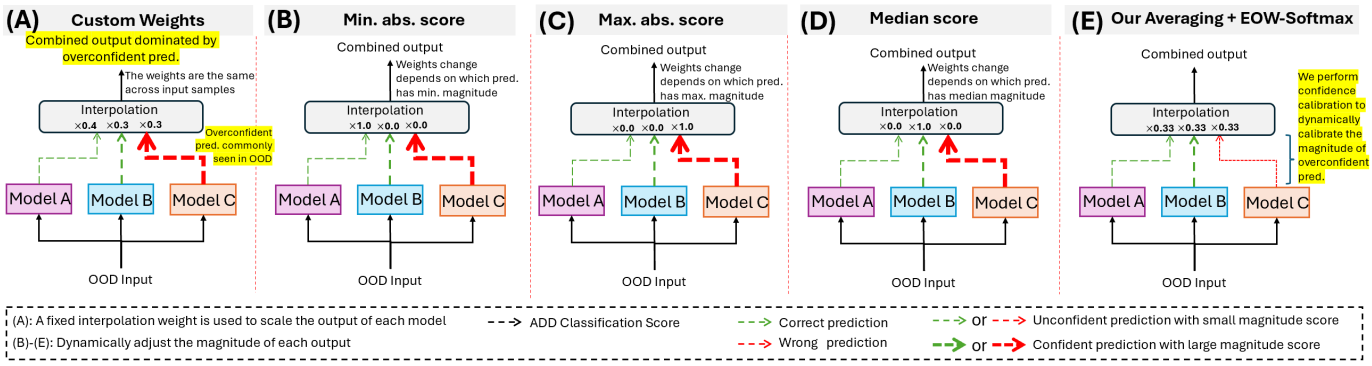

Fig. 2. Overview of our confidence-calibrated model ensemble. As an example, three models A, B, and C are trained to perform ADD in the OOD scenario, and model C produces over-confident and inaccurate predictions due to the OOD input. As ensemble methods based on the likelihood ratio test are not possible for OOD data, previous work combines the model classification scores via linear interpolation based on (A) custom weights, (B) minimum absolute score, (C) maximum absolute score, or (D) median score. However, they rarely explored improving the ensemble output combination using confidence calibration. (E) We find that calibrating the overconfident predictions caused by OOD inputs is crucial in improving the ensemble's output combination due to the inconsistent performance of the models across different subsets of OOD samples, as shown in Fig. 1.

Our contributions are three fold: 1) We show that confidence calibration is an effective approach to adjust the magnitudes of the individual model outputs in a model ensemble to improve the output combination. 2) Our ensemble has significantly improved the equal error rate (EER) from 2.66% to 2.03% on the ASVspoof 2021 deepfake (DF) test set compared to other output combination methods, and 3) we show that EER results from individual models are unstable, and our ensemble method can achieve EER close to the lower bound of the results.

II. Method

A. Baseline XLSR-Conformer

We utilize the XLSR-Conformer [10] architecture as our baseline. XLSR-Conformer is based on the pre-trained XLSR [11] model, a variant of wav2vec 2.0. This architecture benefits from model training on extensive data in an SSL manner, enabling it to extract rich speech representations valuable for numerous tasks [12], including synthetic speech detection [13]. XLSR includes a CNN front-end that converts the 1D raw waveform into a 2D temporal-channel representation and 24 transformer encoder layers to capture global speech relationships. The resulting speech representation has a shape of $(T \times D)$, where T is the temporal length, and D is the channel dimension.

The XLSR representation is then projected to a D-dimensional space and concatenated with a learnable classification token (CLS) to form an input sequence $X \in \mathcal{R}^{(T+1)\times D}$ for the Conformer model. The Conformer model, composed of L Conformer blocks, includes multi-head self-attention (MHSA), feed-forward modules, and additional convolutional layers to capture local dependencies within the speech representation. The CLS token is finally detached from the Conformer model's output to classify the input speech as bona fide or spoof.

B. Energy-based Open World Softmax (EOW Softmax)

To perform confidence calibration on ADD models, the models are trained with EOW Softmax [7]. However, unlike in previous work which uses EOW Softmax to calibrate predicted probabilities, we apply EOW Softmax to improve output combinations for model ensembles.

EOW Softmax models open-world uncertainty as an additional dimension using a $(K+1)$-way softmax formulation, where the first $K = 2$ dimensions correspond to the original bona fide and spoof scores for ADD, and the extra third dimension corresponds to the open-world uncertainty score, where the score is supposed to give a high value when the prediction is of low confidence.

Specifically, let $f_\theta : \mathcal{R}^D \rightarrow \mathcal{R}^{(K+1)}$ be the last linear layer in an ADD model which produces 3 logits denoted as $f_\theta(x)[i]$ with $i \in \{1, 2, 3\}$ given a pooled speech features $x \in \mathcal{R}^D$. The probabilities can be computed from the logits using Softmax normalization:

$$h_\theta(x)[i] = \frac{\exp f_\theta(x)[i]}{\sum_{j=1}^{K+1} \exp f_\theta(x)[j]} \tag{1}$$

where h_θ is the concatenation of the linear layer and a Softmax normalization layer. Then, the loss function $L_{eow}(x, y)$ for EOW Softmax is defined as

$$\begin{aligned} L_{eow}(x, y) = \min_\theta & \mathbb{E}_{p(x)}\Big[-\log h_\theta(x)[y] \Big] \\ & + \lambda \mathbb{E}_{p_{\hat{\theta}}(x)}\Big[-\log h_\theta(x)[K+1] \Big] \end{aligned} \tag{2}$$

where $0 < \lambda$ is a hyper-parameter; the first term is the maximum log-likelihood (MLL) objective for the ADD task using the ground truth label y; the second term is the MLL objective for detecting pooled speech features sampled from normal distributions $p_{\hat{\theta}}(x)$ that estimates $p_\theta(x)$, and the normal distributions are learned using SGLD-based optimizations [8].

Intuitively, $x \sim p_\theta(x)$ represents the pooled speech features computed from real or synthesized audio. $x \sim p_{\hat{\theta}}(x)$ represents the features sampled from normal distributions which may be different from the real or synthesized audio feature

distributions. So, the model should be trained to output a higher open-world uncertainty score for the latter distribution.

C. Confidence-based Model Ensemble

To improve ADD performance, confidence-based model ensembles are used. Specifically, as shown in Figure 2, the scores of multiple models are combined to improve the final prediction of a model ensemble, and we propose to calibrate the confidence of each individual model in the ensemble, such that the improved confidence can be used to control the influence of each individual model's output score on the final combined result. The final score $s(x)$ is then defined as

$$s(x) = \frac{\sum_{j=1}^{N} f_{\theta_j}(x)[1]}{N} \tag{3}$$

where f_{θ_j} is the last linear layer, and $f_{\theta_j}(x)[1]$ is the spoof score output of the j-th model in an ensemble of N models. After confidence calibration, the magnitude of the score output of each individual model will be adjusted according to the confidence of the model prediction, where a larger magnitude means higher confidence. Therefore, scores with higher confidence will influence the final combined score $s(x)$ more due to their higher magnitude.

III. Experiment Setup

A. Dataset and metrics

The training and development data were obtained from the ASVspoof 2019 [14] logical access (LA) track, encompassing clean speech alongside text-to-speech and voice conversion attacks. Our method was assessed using the ASVspoof 2021 [15] DF evaluation set, which comprises bona fide and spoofed speech utterances processed through various lossy codecs typically used for media storage. The data is encoded and then decoded to recover uncompressed audio, introducing distortions that vary based on the codec and its configuration.

B. Implementation details

During training, audio data were cropped or concatenated into segments of approximately 4 seconds (64,600 samples). We used the Adam optimizer with a learning rate of 10^{-6} and a weight decay of 10^{-4} to optimize a weighted cross-entropy loss. The batch size was set to 20. The final result was derived from a model checkpoint created by averaging the top 5 best-performing models on the validation set. Early stopping was implemented if the validation cross-entropy loss did not improve for 7 epochs. EOW Softmax was configured following [8]. All experiments were conducted using a single Nvidia A40 GPU. All models were repeated trained five times with five different random seeds to improve the statistical power of the study.

In terms of model architecture, we utilized the pre-trained SSL model XLSR2 as an upstream model to extract intermediate representations from the raw input signal, following our baseline [10]. To ensure comparability with previous works [10], [13], we applied the RawBoost signal noise injection data augmentation technique [16]. The configuration and parameters of RawBoost in our experiments were consistent with the original paper, where stationary signal-independent additive, randomly colored noise was added during training.

Finally, we refer to our model, which is XLSR-Conformer with EOW Softmax applied to calibrate confidence, as SSL-EOW-Softmax. Our EOW Softmax configuration is the same as Wang et al. [8].

IV. Results and Discussions

A. EOW Softmax for confidence-based ensemble

We show the EER results of model ensembles in Table I, where the ensembles are formed from five XLSR-Conformers [10] trained with different random seeds. All models are trained on the ASVspoof 2019 LA train set and evaluated on the 2021 DF test set. Unless otherwise specified, we assign a custom interpolation weight of 0.2 to each of the five models to form the ensemble output.

The effectiveness of combining individual ADD model outputs to form the final ensemble output is compared between four baseline methods and our method. For the baselines, the worst pooled EER of 3.17% is obtained if the model output score that has the minimum magnitude is chosen, among the five models, as the final output of the ensemble. This is expected as the model that outputs the minimum absolute score usually has low confidence for its prediction, thus using its output as the ensemble output gives more errors.

Then, a better pooled EER of 2.77% is obtained if we choose the median score of the five model output scores as the ensemble output. The EER is further improved to 2.69% if we choose the output score that has the highest magnitude as the ensemble output. This is expected as the model that outputs the maximum absolute score usually has higher confidence for its prediction, thus using its output as the ensemble output gives less errors.

Next, we observe that instead of selecting the output scores with the highest magnitude to form the ensemble outputs, simply averaging the scores also gives a similar pooled EER of 2.66%. We hypothesize that this is because by averaging the model scores, the scores that have a larger magnitude will affect the averaged result more. Thus, averaging has a similar effect compared with choosing the score with the maximum magnitude, as both methods let the maximum-magnitude score influence the final ensemble output more.

Finally, we apply EOW Softmax to all the models in an ensemble to calibrate their confidences and our confidence ensemble method improves the pooled EER significantly from 2.66% to 2.03%. We hypothesis this is because by calibrating the confidence of the models, the model output scores can be better combined.

B. EOW-Softmax for training regularization

In addition to using EOW Softmax to improve the output combination of ensembles as discussed in Section IV-A, EOW Softmax can also be used to reduce overconfident predictions

TABLE I

EER (%) RESULTS OF MODEL ENSEMBLES. THE ENSEMBLES ARE FORMED FROM THE XLSR-CONFORMERS [10] TRAINED FIVE TIMES REPEATEDLY USING DIFFERENT RANDOM SEEDS. ALL MODELS ARE TRAINED ON THE ASVSPOOF 2019 LA TRAIN SET AND EVALUATED ON THE 2021 DF TEST SET. COLUMNS C1 TO C9 SHOWS THE EER OF THE NINE ATTACK TYPES IN THE TEST SET.

Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	Pooled
linear fusion baselines										
min. abs. score	2.95	4.04	3.17	3.32	2.95	2.79	2.48	3.36	2.79	3.17
median score	2.44	3.95	2.73	2.90	2.77	2.31	2.19	2.79	2.44	2.77
max. abs. score	2.51	3.35	2.56	2.79	2.59	2.37	2.05	2.58	2.39	2.69
avg. score	2.42	2.88	2.62	2.71	2.59	2.23	2.13	2.57	2.35	2.66
our method										
conf. ensemble	**1.82**	**2.02**	**1.85**	**2.01**	**1.89**	**1.87**	**1.57**	**2.05**	**1.96**	**2.03**

TABLE II

INDIVIDUAL MODEL EER (%) RESULTS. (*) MEANS OUR REPRODUCED RESULTS, WHERE WE TRAIN THE SAME MODEL FIVE TIMES REPEATEDLY USING DIFFERENT RANDOM SEEDS TO SHOW THE BEST AND AVERAGE EER. OUR ENSEMBLE SINGLE EER RESULT IS SHOWN IN PARENTHESIS.

Method	EER	avg. EER
reported baselines		
Shim et al. [17]	20.84	-
Shim et al. [18]	17.48	-
Rosello et al. [10]	2.27	-
Xie et al. [1]	2.22	-
Truong et al. [19]	2.06	-
reproduced baselines		
Rosello et al. [10]*	2.52	3.50
Truong et al. [19]*	2.33	3.11
our method		
SSL-EOW-Softmax	**1.75**	**2.91**
Calibrated Ensemble		**(2.03)**

TABLE III

ABLATION STUDY ON THE EFFECT OF CONFIDENCE-BASED ENSEMBLE. AS AN UPPER BOUND FOR THE ENSEMBLE EER RESULTS, A MODEL IS RANDOMLY SELECTED FROM AN ENSEMBLE TO PERFORM ADD WITHOUT PERFORMING ANY OUTPUT COMBINATION. THEN, THE EER REDUCTIONS (EERR) FROM THE UPPER BOUND TO THE AVERAGE LINEAR FUSION RESULTS ARE SHOWN.

Method	EER (%)	EERR (%)
baseline 1		
Shim et al. [17]		
w/ random ensem.	3.52	0.0
w/ avg. linear fusion	2.66	24.4
baseline 2		
Truong et al. [19]		
w/ random ensem.	3.11	0.0
w/ avg. linear fusion	2.31	25.7
our method		
SSL-EOW-Softmax		
w/ random ensem.	2.90	0.0
w/ avg. linear fusion	**2.03**	**30.0**

in ADD models to prevent overfitting. Table II shows the results of applying EOW Softmax to an individual model.

As it is known that ADD training is not stable [20], we further reproduce the results of two baseline methods for comparison and report the results of training a model five times repeatedly using different random seeds to improve the statistical power of the study. The results of the reproduced baselines show that the EER reported in previous work is usually lower than the average EER. This suggested that ADD training is sometimes not stable and may give higher EER. Finally, our method which combines SSL-based ADD models with EOW Softmax gives the best EER result compared to previous work.

C. Ablation Study

As individual models trained with EOW Softmax have improved performance, it may be unclear whether the improvements in the EOW-Softmax-trained model ensemble in Table I are caused by the improvement in performance of the individual model or are caused by the improved output combination after confidence calibration. To verify, Table III shows the difference in EER reduction (EERR) when using the confidence-calibrated ensemble versus the uncalibrated ensemble. The results show that our ensemble of confidence-calibrated models achieves a larger EERR of 30.0% compared to uncalibrated ensembles, which achieves an EERR of 24.4% and 25.7%. This shows that model ensembles trained with EOW Softmax can better combine the outputs from multiple models to achieve a larger EERR.

V. CONCLUSION

In conclusion, This study demonstrates the effectiveness of confidence calibration in enhancing ADD ensemble performance in OOD scenarios, reducing the equal error rate from 2.66% to 2.03%. Ablation studies validate the approach, highlighting the importance of addressing model confidence for improved output combination and overall performance. Future work could explore applying this technique to other speech processing tasks and more challenging OOD scenarios.

ACKNOWLEDGMENT

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

REFERENCES

[1] Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye, "Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection," in *Proc. INTERSPEECH*, 2023, vol. 2023, pp. 2808–2812.

[2] Emanuela Marasco and Carlo Sansone, "Improving the accuracy of a score fusion approach based on likelihood ratio in multimodal biometric systems," in *Image Analysis and Processing–ICIAP 2009: 15th International Conference Vietri sul Mare, Italy, September 8-11, 2009 Proceedings 15*. Springer, 2009, pp. 509–518.

[3] Bhusan Chettri, Daniel Stoller, Veronica Morfi, Marco A Martínez Ramírez, Emmanouil Benetos, and Bob L Sturm, "Ensemble models for spoofing detection in automatic speaker verification," *arXiv preprint arXiv:1904.04589*, 2019.

[4] Yuxiang Zhang, Jingze Lu, Xingming Wang, Zhuo Li, Runqiu Xiao, Wenchao Wang, Ming Li, and Pengyuan Zhang, "Deepfake detection system for the add challenge track 3.2 based on score fusion," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022, pp. 43–52.

[5] Mingxing He, Shi-Jinn Horng, Pingzhi Fan, Ray-Shine Run, Rong-Jian Chen, Jui-Lin Lai, Muhammad Khurram Khan, and Kevin Octavius Sentosa, "Performance evaluation of score level fusion in multimodal biometric systems," *Pattern Recognition*, vol. 43, no. 5, pp. 1789–1800, 2010.

[6] Zheng Wang, Sanshuai Cui, Xiangui Kang, Wei Sun, and Zhonghua Li, "Densely connected convolutional network for audio spoofing detection," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 1352–1360.

[7] Yuxiang Zhang, Jingze Lu, Zhuo Li, Zengqiang Shang, Wenchao Wang, and Pengyuan Zhang, "Improving the robustness of deepfake audio detection through confidence calibration.," in *DADA@ IJCAI*, 2023, pp. 70–75.

[8] Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li, "Energy-based open-world uncertainty modeling for confidence calibration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9302–9311.

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[10] Eros Roselló Casado, Alejandro Gómez Alanís, Ángel Manuel Gómez García, Antonio Miguel Peinado Herreros, et al., "A conformer-based classifier for variable-length utterance processing in anti-spoofing," 2023.

[11] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al., "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[12] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6989–6993.

[13] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233*, 2022.

[14] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[15] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al., "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.

[16] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6382–6386.

[17] Hye-jin Shim, Jee-weon Jung, and Tomi Kinnunen, "Multi-dataset co-training with sharpness-aware optimization for audio anti-spoofing," *arXiv preprint arXiv:2305.19953*, 2023.

[18] Hye-jin Shim, Md Sahidullah, Jee-weon Jung, Shinji Watanabe, and Tomi Kinnunen, "Beyond silence: Bias analysis through loss and asymmetric approach in audio anti-spoofing," *arXiv preprint arXiv:2406.17246*, 2024.

[19] Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng, "Temporal-channel modeling in multi-head self-attention for synthetic speech detection," *arXiv preprint arXiv:2406.17376*, 2024.

[20] Xin Wang and Junich Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," *arXiv preprint arXiv:2103.11326*, 2021.