



Audio Deepfake Source Tracing using Multi-Attribute Open-Set Identification and Verification

Pierre Falez¹, Tony Marteau¹, Damien Lolive², Arnaud Delhay³

¹Whispeak, France

²Univ Bretagne Sud, CNRS, IRISA, France

³Univ Rennes, CNRS, IRISA, France

{pfalez,tmarteau}@whispeak.io,firstname.lastname@irisa.fr

Abstract

Audio deepfake detection has advanced significantly, in particular, thanks to the ASVSpooof challenge. However, existing approaches primarily rely on binary classification, which does not provide information about the origin of manipulated audio. In this paper, we address the problem of source tracing and propose two protocols to evaluate model performance in an open-set setting: (1) a few-shot identification protocol, where K reference audios are provided, and (2) a verification protocol inspired by speaker verification. We classify either the entire generation system or its components, such as the acoustic model or vocoder. Our models are trained both on an internal dataset and on the MLAAD source tracing dataset. Evaluation is done on five public datasets: three ASVSpooof sets, MLAAD and Blizzard23. Results show promising discrimination of unseen class attributes. Finally, we emphasize the need for a standardized ontology for source tracing in audio deepfake detection.

Index Terms: Source Tracing, Antispoofing, Audio Deepfake Detection

1. Introduction

Audio deepfake generation made huge progress in the last few years [1, 2, 3] reducing drastically the perception gap between synthetic and natural speech to a point where the difference is barely noticeable. As a consequence, audio deepfake detection gained in popularity and became a major challenge. In this domain, ASVSpooof challenges focus on the improvement of the detection performance [4, 5, 6].

Although detecting if an audio sample is a spoof or bonafide is crucial, determining which components have been used to generate a fake audio sample is a crucial task. This task needs to move from a binary classification task, telling if a sample is a spoof or not, to a multi-attribute classification task. The latter consists of predicting the different building blocks of the generation system in an independent manner. Furthermore, new components may be created after the training phase of the system, hence making the problem as an open-set classification problem.

Some recent works have focused on identifying the origin of audio deepfake [7], by classifying multiple attributes of spoof systems. Nevertheless, still too few approaches are suggested and the evaluation framework is not yet well-defined.

More precisely, [7] uses two approaches: an end-to-end and a two-stage. End-to-end approach consist to train a classifier for each attribute. Two-stage approach first trains a model following binary spoof classification, then the classification layer is replaced by classification head for each attribute. However, their work is limited to the scope of closed-set classification, where all possible systems are known during training. Unseen classes are not considered, leaving a critical gap in real-world applicability.

In this work, we extend previous work by focusing on open-set classification. We introduce two evaluation strategies: a few-shot learning approach and a verification-based approach. Additionally, we compare various models on this task, leveraging transfer learning to adapt models trained for related tasks, such as spectrogram reconstruction or speaker verification.

The remainder of the paper is organized as follows. Evaluation protocols are presented in detail in section 2. All the datasets used for this study are introduced in section 3. Then, sections 4 and 5 give the details about the models and the training parameters. Finally, section 6 presents the results.

2. Evaluation Protocols

In this work, we consider the source tracing task as an open-set multi-attribute classification problem. Several mutually exclusive classes can be inferred, such as the different building blocks of deepfake generation systems (e.g., vocoder and acoustic model), the datasets used to train them, and the input type (text or voice). In this work, we consider only three attributes: the system as a whole, the acoustic model and the vocoder.

During evaluation, some classes remain unseen (i.e., not used during training), making this an open-set problem. To tackle this challenge, we employ continuous embedding spaces for each attribute. A similarity function s is used to compute the distance between two samples. Data belonging to the same class should be close in the embedding space while being distant from samples of other classes.

In this work, we explore two evaluation protocols to measure the performance of our systems: A few-shot identification protocol and a verification of trials protocol. For the few-shot protocol we separate the evaluation dataset into two splits: one containing a few examples per class which aim to create a reference and the second split containing the remaining of the samples to test the system performance against the references. To create the splits, we randomly select K (5 in this paper) samples for each class in the test dataset. All the remaining samples are put in the second split. References are then created by averaging the K embeddings of selected samples. Lastly, the evaluation is performed following a traditional identification task: the sample's class is predicted by finding the most similar reference.

The second approach to evaluate our systems is directly inspired by speaker verification. For each dataset, we generate a list of target and non-target trials. For each sample in the dataset, we randomly select K_{tgt} samples that belong to the same class, and K_{ntgt} samples that belong to a different class. K_{tgt} and K_{ntgt} are chosen for computational reasons. Values used in this work are reported in Table 2. Thus, this protocol can be treated as a binary classification problem: finding the best threshold that separates targets from non-targets.

The choice of the testing protocol depends on the goal. Few-shot learning is applicable if some reference samples are

Table 1: Training datasets used in this work.

		MLAAD [9]	Whispeak
Train	samples	39 000	471 852
	systems	19	357
Dev	samples	40 000	7 424
	systems	18	256

Table 2: Testing datasets used in this work. Only spoof samples are used. K_{tgt} and K_{ntgt} stand for the number of samples randomly selected for evaluation resp. for the target class and the non-target class.

Dataset	Samples	Systems	K_{tgt} / K_{ntgt}
MLAAD (test) [9]	114 000	41	1 / 5
ASVSpooF19 LA (test) [4]	63 882	13	1 / 5
ASVSpooF21 DF (eval) [5]	519 059	110	1 / 1
ASVSpooF5 (eval) [6]	542 086	16	1 / 1
Blizzard23 (all) [12]	39 556	20	1 / 5

available for each class. Verification is used when the goal is to determine whether two samples originate from a similar class.

3. Datasets

As it is a recent task, few public datasets are available for audio deepfake source tracing. For our study we have therefore chosen to work with the MLAAD dataset, with a larger internal dataset, and 4 public evaluation datasets hijacked for this task.

MLAAD [8] It is a multilingual dataset that gathers samples generated from 82 text-to-speech systems. We use the version 5 in this work. The samples are split into training, development and testing sets following the source tracing protocol [9].

Our internal dataset (called WHISPEAK) is constructed by combining synthetic audio from several open-source models (ESPNet [10], CoquiTTS [11], ...) and commercial models. As a multilingual dataset, it covers 7 languages (English, French, Spanish, German, Russian, Turkish and Arabic) with 357 distinct systems. Systems are waveform concatenation, self-vocoded, text-to-speech and speech-to-speech technologies of predefined voices and cloned voices. In this dataset, each sample is labeled with multiple attributes corresponding to the acoustic model (21 classes, e.g., vits, tacotron2, transformer-tts, ...), the vocoder type (10 classes, e.g., hifigan, wavegrad, melgan, ...) and the system. Samples are distributed into training and development sets.

Table 1 sums up the number of samples and the number of systems in each set for these two training datasets.

Our evaluations are done on the MLAAD source tracing testing set and four other public datasets derived for this task: three ASVSpooF [4, 5, 6] datasets and the Blizzard23 dataset [12]. For all of these datasets only fake audio samples are used, bonafide samples are removed. These datasets are described in terms of number of samples and number of systems in the Table 2. These datasets cover multiple years, systems and quality of speech synthesis generations.

For the Blizzard23 dataset, we have relabeled the acoustic model and the vocoder with information provided by authors in their presentation paper. From 20 systems in this dataset, 11 acoustic models and 7 vocoder types are present. This dataset is the only dataset used to evaluate the multiple attributes predictions.

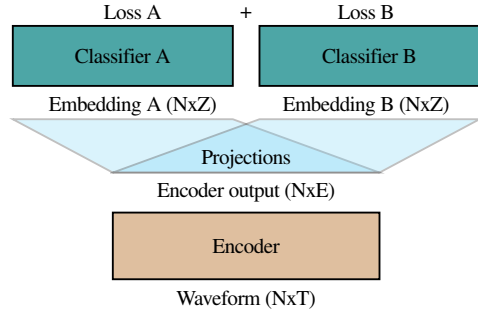


Figure 1: Schema of the model. An encoder extracts features from the waveform. Features are projected into embedding spaces for each attribute. During training, for each attribute a classifier is added on top of each embedding.

4. Models

The model used in this work consists of multiple building blocks (See Figure 1). The core component is the encoder, which converts a waveform of shape $(N \times T)$ into a time-independent representation $(N \times E)$. A projection layer is added on top of the encoder for each target attribute (e.g., system, acoustic model, vocoder), producing embeddings of shape $(N \times Z)$, where Z is fixed at 64 in this work. During training, a classifier is appended following each projection layer to classify the corresponding attribute of training data. However, for the evaluation, we don't use the classifiers. Instead, we use a cosine metric in order to compute the similarity of two samples in the embedding spaces.

Recent work suggests that the use of transfer learning is essential to improve the generalization ability of models on audio deepfake detection tasks [13, 8]. Based on this observation, we evaluate three types of pre-trained models as encoder blocks: Wav2Vec [14], commonly used in the deepfake detection community, ResNet-293 trained for speaker verification [15], and M2D [16] trained to reconstruct masked part of the input. M2D and ResNet-293 are models based on a spectrogram input while Wav2Vec is a model based on the raw waveform. Each model extracts different features as they have been pre-trained on different tasks.

Models trained on MLAAD dataset only use the whole system as attribute (the field architecture in the metadata). Models trained on WHISPEAK dataset use four attributes: system name, system type (TTS, S2S, Waveform Concat, Self Vocoded), acoustic model and vocoder.

5. Training Details

Although pre-trained encoders are utilized, all models are fine-tuned in an end-to-end manner for 200k updates with the ADAM optimizer using a learning rate of $1e^{-6}$ in order to avoid overfitting. We use a weight decay of $1e^{-4}$. Batches are created by using random chunks of 4 seconds of audio (6 seconds for the M2D model). If audio is smaller, a circular padding is employed. We use multiple data augmentation techniques during the training process with a probability $p_{DA} = 0.2$ to be used for each sample (More details can be found in [17]). We use silence removal, noise addition from MUSAN [18], reverberation from RIRS.NOISES, Rawboost [19], codec applications and SpecAugment [20].

We use a batch size of 128 for the M2D and Wav2vec models, and 64 for the ResNet-293 for memory reasons. For the AAM-Softmax loss, we set the scale to 30 and the margin to 0.2. All the models are trained on a cluster of L40S.

6. Results

This section presents results based on the two protocols detailed previously. Inference is performed on the first 4 seconds of each audio sample, except for the M2D model, which uses 6 seconds. If audio sample is smaller, a circular padding is employed. In the following, the Top-3 accuracy is a relaxed version of accuracy where one prediction is considered to be correct if it is present in the first three ground truth. Top-1 accuracy is the standard macro averaged accuracy. Finally, the Equal-Error Rate (EER) is also used to compare the systems.

6.1. Few-shot Identification Protocol

Table 3 shows Top-1 and Top-3 accuracy to predict the system when using the few-shot protocol with $K = 5$ references for each system. These results are reported on selected test datasets for models trained on either the MLAAD dataset or our internal WHISPEAK dataset.

The first conclusion from these results is that models trained on our internal WHISPEAK dataset outperform those trained on MLAAD on all selected test datasets. Several factors may contribute to this improvement. One possibility is that our internal dataset encompasses a greater diversity of systems, leading to more precise embeddings. Additionally, the multi-head classification used during training with the WHISPEAK dataset may help constrain the projection embedding within the system head, further enhancing performance.

Another key observation is that, when evaluated on the MLAAD test dataset, the performance gap between models trained on MLAAD and those trained on WHISPEAK is smaller. However, this gap widens significantly when tested on ASVSpooof19 LA, ASVSpooof21 DF, ASVSpooof5 and Blizzard23.

Additionally, even if the better Top1 accuracy is low, especially for the ASVSpooof21 DF and ASVSpooof5, the Top3 accuracy is much better. This poor performance can be explained by the larger number of systems in the datasets and a confusion on the embedding of similar systems.

We analyze classes with most confusion for the M2D model trained on WHISPEAK dataset on the MLAAD evaluation set (see Figure 2). Most systems are well classified. However some of them, like *bark*, *parler_tts_mini_v1* and *suno/bark-small* have high error rates, respectively 87.62%, 72.26% and 71.36%. Those classes are mainly confused with similar models (*bark* with *suno/bark* and *suno/bark-small*, *parler_tts_mini_v1* with *parler_tts_large_v1*). This highlights that similar models appear to be close in the embedding space. This also underlines the need for hierarchical model classification.

We also looked at the accuracy of seen and unseen class. For the M2D model trained on MLAAD, we have a top-1 accuracy of 86.34% for known classes and 62.38% for unknown. This shows that the model is less accurate on unseen classes, but is still able to generalize.

With models trained on WHISPEAK dataset we also evaluate other head on the Blizzard23 dataset (see Table 4). Contrary to the observation made in [21], it seems more complicated to classify vocoders than acoustic models.

We can observe that classifying the whole system seems easier than classifying the building blocks. We also compared two ways of classifying the system: by considering it as a whole, or by the combination of the different building blocks, by concatenating the embeddings. Our results show that classifying the system as a whole gives better performances. This may be due to annotation errors of building blocks in the train set.

For a more in-depth analysis, the Venn diagram (Figure 3) highlights the overlap in accurate identification between heads with

the ResNet-293 models trained on WHISPEAK dataset and tested on Blizzard23 dataset. We can confirm classifying the building blocks is a more difficult task (29782 good system identifications vs. 16520 and 22062 good identifications for vocoder and acoustic model). Furthermore, lots of good building blocks identifications are common with system identifications (4694, 9145 and 8209). Next, we can observe that identifying the acoustic model is easier than identifying the vocoder (22062 vs. 16520 good identifications). And lastly the Venn representation confirms the use of vocoder and acoustic heads to identify the system is less efficient than identifying the system directly.

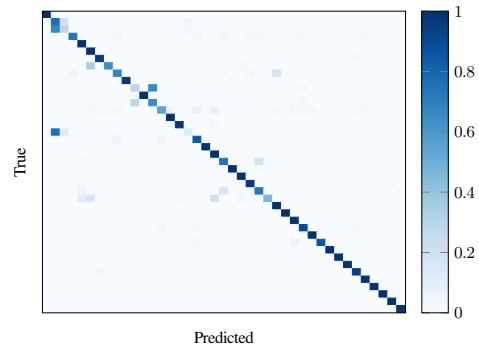


Figure 2: Confusion matrix for the M2D model trained on WHISPEAK and tested on the MLAAD dataset.

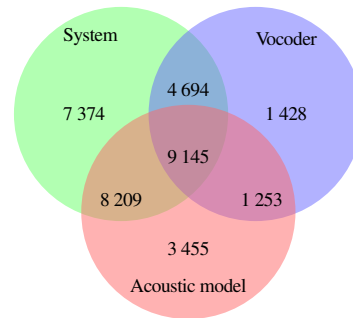


Figure 3: Venn representation of good identifications of system, vocoder and acoustic model with ResNet-293 trained on WHISPEAK dataset with the few-shot protocol. For instance, 8209 times the system and the acoustic model have been identified correctly together.

6.2. Verification Protocol

The EER results for the verification protocol are presented in Table 5. The key trends observed in the identification protocol also hold in this setting. However, one notable difference is that the ASVSpooof5 dataset appears more challenging than ASVSpooof21, primarily because its results are less influenced by the number of classes.

A closer examination of the most frequently confused systems (Table 6) reveals a different set of systems compared to the few-shot protocol (*bark*, *parler_tts_mini_v1*, and *suno/bark-small* vs. *tacotron-DDC*, *glow-tts*, *vits*, etc.). To understand this difference, we analyzed the occurrence frequency of each class and found that these confused systems are not among the most represented in the dataset. This finding supports the idea that using multiple samples can help generate more accurate embeddings. A second

Table 3: Systems identification accuracy (%) \uparrow on selected test datasets using Few-shot protocol.

Model	Training	MLAAD		ASVSpooof19 LA		ASVSpooof21 DF		ASVSpooof5		Blizzard23	
		Top-1	Top-3	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3
M2D	MLAAD	72.32	87.81	52.79	78.45	14.48	28.78	29.89	57.74	48.79	72.64
	WHISPEAK	85.28	97.47	85.93	97.26	42.16	67.34	54.97	77.33	72.33	88.66
Wav2Vec	MLAAD	79.27	92.04	65.65	87.44	24.87	43.90	43.40	74.41	39.49	62.11
	WHISPEAK	78.26	92.80	88.36	98.71	47.72	75.17	61.92	84.46	50.46	70.76
ResNet-293	MLAAD	83.57	93.09	61.35	86.98	22.02	39.28	35.60	65.29	53.66	72.58
	WHISPEAK	84.18	96.37	83.06	97.73	32.37	56.02	43.94	70.12	74.94	88.33

Table 4: Attribute identification accuracy (%) \uparrow on Blizzard23 with models trained on WHISPEAK using Few-shot protocol.

Model	System		Acoustic Model		Vocoder		Acoustic Model + Vocoder	
	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3
M2D	72.33	88.66	63.21	83.23	49.88	79.96	64.47	84.31
Wav2Vec	50.46	70.76	57.91	81.50	53.44	82.27	52.72	73.93
ResNet-293	74.94	88.33	71.49	87.00	63.55	85.22	69.26	83.31

Table 5: Models EER (%) \downarrow on selected test datasets using verification protocol.

Model	Training	MLAAD	ASVSpooof19 LA	ASVSpooof21 DF	ASVSpooof5	Blizzard23
M2D	MLAAD	21.87	28.93	35.26	36.07	24.79
	WHISPEAK	17.95	15.14	20.34	31.34	18.44
Wav2Vec	MLAAD	19.64	22.93	28.03	32.02	28.36
	WHISPEAK	25.88	14.03	15.12	24.85	22.53
ResNet-293	MLAAD	21.26	27.64	32.85	32.66	23.67
	WHISPEAK	20.78	13.02	23.44	34.15	18.10

Table 6: Classes with higher EER (%) on the MLAAD dataset with the M2D model.

Class	EER	Class	EER
tacotron2-DDC	52.87	facebook/mms-tts-hun	21.88
glow-tts	39.02	Mars5	21.20
vits	37.47	OpenVoiceV2	20.66
tacotron2-DCA	26.36	e2-tts	20.20
griffin_lim	22.90	f5-tts	19.70

explanation could be that in the verification protocol, all the systems are compared with a single threshold, whereas in the few-shot protocol, the closest reference embedding is used.

7. Conclusion and Future Works

Source tracing of deepfake audio is an emerging task, with only a few studies on the subject. Early work limited themselves to closed classification of system attributes. We have shown that it is possible to extend this task to open-set classification, by using the few-shot or the verification protocol. First results are encouraging, showing satisfactory results for different datasets. We have also compared the use of different training datasets, showing that it is necessary to increase their size to enhance performance. Finally, we compare different encoders and show their relevance regarding the evaluation dataset.

However, that there is currently no agreement on the different attributes that can be used for source tracing tasks, nor are there any on the classes of each attribute. For example, it is quite difficult

to make a list of possible acoustic models or vocoders. Ideally, a hierarchy should be created to represent the similarity of classes. The few works on source tracing arbitrarily label different datasets. In order to solve these problems, we feel it is important to define a common ontology to better structure the future work around source tracing of deepfake audio.

8. References

- [1] T. M. Wani, S. A. A. Qadri, F. A. Wani, I. Amerini *et al.*, “Navigating the soundscape of deception: A comprehensive survey on audio deepfake generation, detection, and future horizons,” *Foundations and Trends® in Privacy and Security*, vol. 6, no. 3-4, pp. 153–345, 2024.
- [2] Y. Xie, X. Wang, Z. Wang, R. Fu, Z. Wen, S. Cao, L. Ma, C. Li, H. Cheng, and L. Ye, “Neural codec source tracing: Toward comprehensive attribution in open-set condition,” *arXiv preprint arXiv:2501.06514*, 2025.
- [3] M. Chhibber, J. Mishra, H. Shim, and T. H. Kinnunen, “An explainable probabilistic attribute embedding approach for spoofed speech characterization,” *arXiv preprint arXiv:2409.11027*, 2024.
- [4] 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.
- [5] 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,” in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.
- [6] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen *et al.*, “Asvspoof 5:

- Crowdsourced speech data, deepfakes, and adversarial attacks at scale,” *arXiv preprint arXiv:2408.08739*, 2024.
- [7] N. Klein, T. Chen, H. Tak, R. Casal, and E. Khoury, “Source tracing of audio deepfake systems,” *arXiv preprint arXiv:2407.08016*, 2024.
- [8] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, “MLAAD: the multi-language audio anti-spoofing dataset,” in *International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, June 30 - July 5*. IEEE, 2024, pp. 1–7.
- [9] N. Müller, “Using mlaad for source tracing of audio deepfakes,” <https://deepfake-total.com/sourcetracing>, Fraunhofer AISEC, 11 2024.
- [10] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “Espnet: End-to-end speech processing toolkit,” in *International Speech Conference (INTERSPEECH), Hyderabad, India, September 2-6*. ISCA, 2018, pp. 2207–2211.
- [11] G. Eren and The Coqui TTS Team, “Coqui TTS,” 2021. [Online]. Available: <https://github.com/coqui-ai/TTS>
- [12] O. Perrotin, B. Stephenson, S. Gerber, and G. Bailly, “The blizzard challenge 2023,” in *18th Blizzard Challenge Workshop*. ISCA, 2023, pp. 1–27.
- [13] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, “Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,” in *The Speaker and Language Recognition Workshop (Odyssey), Beijing, China, 28 June - 1 July 2022*. ISCA, 2022, pp. 112–119.
- [14] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, “XLS-R: self-supervised cross-lingual speech representation learning at scale,” in *International Speech Conference (INTERSPEECH), Incheon, Korea, September 18-22*. ISCA, 2022, pp. 2278–2282.
- [15] Y. Lin, M. Cheng, F. Zhang, Y. Gao, S. Zhang, and M. Li, “Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark,” *arXiv preprint arXiv:2407.11510*, 2024.
- [16] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, M. Yasuda, S. Tsubaki, and K. Imoto, “M2d-clap: Masked modeling duo meets clap for learning general-purpose audio-language representation,” *arXiv preprint arXiv:2406.02032*, 2024.
- [17] P. Falez and T. Marteau, “Whispeak speech deepfake detection systems for the asvspoof5 challenge,” in *Proc. ASVspoof 2024*, 2024, pp. 32–35.
- [18] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,” 2015, arXiv:1510.08484v1.
- [19] H. Tak, M. R. Kamble, J. Patino, M. Todisco, and N. W. D. Evans, “Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022, Virtual and Singapore, 23-27 May*. IEEE, 2022, pp. 6382–6386.
- [20] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” in *International Speech Conference (INTERSPEECH), Graz, Austria, September 15-19*. ISCA, 2019, pp. 2613–2617.
- [21] C. Y. Zhang, J. Yi, J. Tao, C. Wang, and X. Yan, “Distinguishing neural speech synthesis models through fingerprints in speech waveforms,” in *China National Conference on Chinese Computational Linguistics*. Springer, 2024, pp. 259–273.