

NE-PADD: Leveraging Named Entity Knowledge for Robust Partial Audio Deepfake Detection via Attention Aggregation

Huhong Xian*, Rui Liu*[†], Berrak Sisman[‡] and Haizhou Li[§]

* Inner Mongolia University, Hohhot, China

E-mail: xianhuhong@163.com; liurui_imu@163.com

[‡] Center for Language and Speech Processing (CLSP), Johns Hopkins University, USA

E-mail: sisman@jhu.edu

[§] School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China

E-mail: haizhouli@cuhk.edu.cn

Abstract—Different from traditional sentence-level audio deepfake detection (ADD), partial audio deepfake detection (PADD) requires frame-level positioning of the location of fake speech. While some progress has been made in this area, leveraging semantic information from audio, especially named entities, remains an underexplored aspect. To this end, we propose a novel method, NE-PADD, which leverages named entity knowledge for robust partial audio deepfake detection through attention aggregation. NE-PADD consists of two parallel branches: Speech Named Entity Recognition (SpeechNER) and PADD. Specifically, we introduce two attention aggregation mechanisms to help PADD models better understand named entity knowledge, thus achieving more robust PADD performance. (1) Attention Fusion (AF) combines attention from SpeechNER and PADD for more accurate weights. (2) Attention Transfer (AT) uses an auxiliary loss from their attention distributions to guide PADD in learning named entity semantics. We construct an appropriate PartialSpoofNER dataset based on the existing PartialSpoof dataset and conduct a detailed comparative analysis of the two attention aggregation methods. Experiments demonstrate that our method exceeds all advanced baselines and demonstrates the effectiveness of fusing named entity knowledge during PADD. The code is available at <https://github.com/AI-S2-Lab/NE-PADD>.

I. INTRODUCTION

In recent years, deep learning has catalyzed remarkable advancements in audio spoofing technologies, particularly in the domains of text-to-speech (TTS) [1]–[3] and voice conversion (VC) [4]–[7]. By leveraging generative models such as CosyVoice¹ and GPT-SoVITS², TTS and VC systems now produce synthetic speech with near-human quality, rendering it nearly indistinguishable from genuine audio [8]–[11]. Partially spoofed audio refers to the insertion or splicing of synthesized speech segments into authentic audio, posing significant risks. Attackers manipulate small, specific units (such as words, characters, or even phonemes) to alter the meaning of sentences, thereby deceiving both machines and humans. Effectively

detecting and mitigating such spoofed audio has become a critical and urgent challenge.

Currently, numerous methods and datasets have been developed for partial audio deepfake detection (PADD). RawNet2 [12] processes raw audio using Sinc-Layers and residual blocks for efficient partial audio deepfake detection. AASIST [13] combines a RawNet2-based encoder and RawGAT-ST for integrating temporal and spectral features using graph attention layers. Yi et al. [14] created the first partially spoofed audio dataset, HAD, in which certain natural speech segments are replaced with synthesized speech segments that carry different semantic content. Zhang et al. [15] introduced the PartialSpoof dataset, designed to provide segment-level labels for audio spoofing detection. Building on this dataset, they explored a multi-task learning framework capable of identifying spoofing occurrences at both the segment and utterance levels [16]. Subsequently, they further extended this work, and Wav2vec2 [17] was adopted as a front-end feature extractor to enable simultaneous detection of both segment-level and utterance-level spoofing [18].

Although existing methods have made significant progress in PADD, the spoofed segments often appear in parts of the audio containing named entities. While previous PADD methods have demonstrated some effectiveness, they often overlook semantic information, particularly the role of named entities. In fact, named entities play a critical role in PADD. For example, partial deepfakes in everyday language frequently involve manipulating one or more named entities within a sentence. Consider the sentence: “*CompanyX CEO heads a meeting in their Shenzhen office.*” Here, “CompanyX” refers to an organization and “Shenzhen” is a location. If any of these entities were replaced or altered, a partial deepfake could occur, which is often difficult to detect using traditional methods. Therefore, changes to named entities are crucial in PADD. Furthermore, some partial spoofing datasets (such as HAD) are specifically designed by replacing named entities, further emphasizing the central role of named entities in partial

[†] Corresponding author.

¹<https://github.com/FunAudioLLM/CosyVoice>

²<https://github.com/RVC-Boss/GPT-SoVITS>

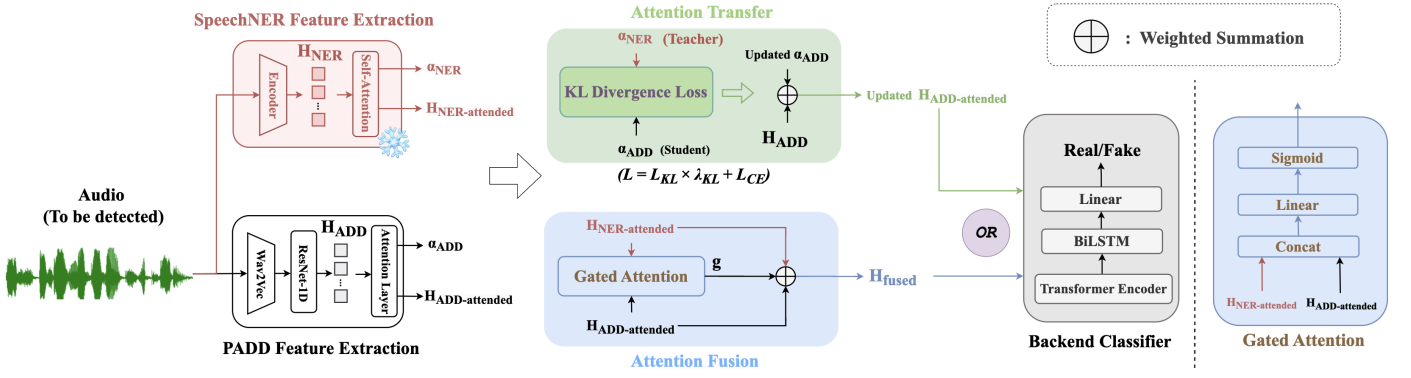


Fig. 1. The architecture of our proposed model. The model architecture includes two feature extraction modules on the left, with the proposed attention fusion (AF) and attention transfer (AT) methods in the center, and the backend classifier on the right.

deepfake tasks. As such, exploring ways to leverage named entity information to improve PADD is an important research direction that warrants further investigation.

To address these challenges, we propose NE-PADD, a novel PADD framework that enhances detection by integrating semantic information from audio named entities using two attention mechanisms. We also introduce PartialSpoofNER, a dataset based on PartialSpoof enriched with named entity annotations. Extensive experiments show that NE-PADD outperforms existing methods on key metrics such as Equal Error Rate, demonstrating its effectiveness and setting a new benchmark in partial audio deepfake detection. The main contributions are summarized as follows:

- We propose a novel model called NE-PADD, which integrates the semantic information of named entities in audio into the deepfake detection model using an attention mechanism.
- We design an approach that integrates SpeechNER and PADD feature extraction modules to learn attention weights and extract intermediate features for spoof detection. We use Attention Fusion and Transfer to extract semantic information from speech named entities, ultimately improving the detection performance.
- We develop a new dataset called PartialSpoofNER, based on the PartialSpoof dataset and enriched with named entity annotations, to provide semantic cues for spoofed audio detection.

II. PROPOSED METHOD

A. Problem Statement and Overview

In partially spoofed scenarios, spoofed audio segments are embedded within authentic speech. Our goal is to detect these spoofed segments at the frame-level. Given the large-scale self-supervised acoustic features:

$$\mathbf{X} = \{f_1, f_2, \dots, f_T\} \in \mathbb{R}^{D \times T}, \quad (1)$$

where D and T represent the feature dimensions and the number of frames, respectively. The task is framed as a binary classification problem at the frame-level, where for each input feature \mathbf{X} , the output is a sequence of frame-level labels:

$$\mathbf{y} = \{y_1, y_2, \dots, y_T\} \in \{0, 1\}^T. \quad (2)$$

Here, $y_t = 1$ indicates an authentic frame, while $y_t = 0$ represents a spoofed frame. Since spoofed segments are often subtle and localized, achieving precise boundary detection is particularly challenging.

In our proposed model, the attention fusion method generates gated attention weights through the Gated Attention module, which are then applied to the two distinct intermediate features, $\mathbf{H}_{\text{ADD-attended}}$ and $\mathbf{H}_{\text{NER-attended}}$, for weighted summation, producing a new fused feature $\mathbf{H}_{\text{fused}}$. The $\mathbf{H}_{\text{fused}}$ is subsequently passed to the frame-level classifier for the final detection of spoofed segments. The attention transfer method computes the Kullback–Leibler (KL) divergence between the two attention distributions, α_{ADD} and α_{NER} , and incorporates it as an auxiliary loss, weighted by a specific factor into the main loss, effectively integrating the semantic information of named entities. The architecture of the proposed model is shown in Fig. 1.

B. PADD Feature Extraction

We utilize the pretrained Wav2Vec2.0 model, trained on the 960-hour LibriSpeech corpus, to extract frame-level features with a frame rate of 20ms [19]. The features are further processed by ResNet-1D, which consists of multiple residual blocks. These blocks effectively capture temporal dependencies and local patterns in the audio, outputting embeddings (\mathbf{H}_{ADD}) that serve as input to the attention mechanisms.

Next, the \mathbf{H}_{ADD} are passed through an attention layer to dynamically focus on relevant regions of the audio. The attention scores are calculated as:

$$\alpha_{\text{ADD}} = \text{softmax} \left(\frac{\mathbf{Q}_{\text{ADD}} \mathbf{K}_{\text{ADD}}^{\top}}{\sqrt{D}} \right). \quad (3)$$

Here, α_{ADD} represents the dynamic attention weights that highlight potentially spoofed segments in the audio. The attended embeddings ($\mathbf{H}_{\text{ADD-attended}}$) are computed as:

$$\mathbf{H}_{\text{ADD-attended}} = \alpha_{\text{ADD}} \mathbf{V}_{\text{ADD}}. \quad (4)$$

The attention-modified embeddings ($\mathbf{H}_{\text{ADD-attended}}$) provide a focused representation of the audio input, emphasizing regions that are likely to contain spoofed segments.

C. SpeechNER Feature Extraction

The SpeechNER feature extraction module identifies named entities by adding special symbols before and after the named entities in the transcribed text of the input audio. Given the scarcity of research on Chinese speech named entities, our experiments are currently limited to English datasets. We draw inspiration from the work of Yadav et al. [20] on end-to-end named entity recognition from English speech, focusing on recognizing the three most frequent named entities (organization, person, and location) from English speech datasets. For this purpose, we use three special symbols ('{', '|', '\$') to denote the start of different named entities, with ']' serving as a common symbol to mark the end of all named entities. As illustrated by the audio example proposed in the Introduction section of our paper: "*CompanyX CEO heads a meeting in their Shenzhen office.*" The actual training text label appears as "{CompanyX} CEO heads a meeting in their \$Shenzhen] office," where '{' and '\$' denote the start of named entities, and ']' marks their end. This is treated as a sequence labeling task, where the input audio is transformed into log-spectrograms of power-normalized audio clips using a 20 ms Hamming window [20], following the common practice in speech processing.

The SpeechNER feature extraction module first processes the named entity information from the audio input through an encoder. The encoder handles audio features, including convolutional and normalization layers, followed by multiple bidirectional long short-term memory (BiLSTM) layers to capture temporal dependencies in the audio data. This process generates named entity representations (\mathbf{H}_{NER}) suitable for further processing, encapsulating contextual information at each time step of the audio.

Subsequently, a self-attention mechanism is applied to process these representations. The attention mechanism computes attention scores to generate entity-level embeddings. These attention-modified embeddings ($\mathbf{H}_{\text{NER-attended}}$) highlight the significance of named entities, providing crucial semantic information for the audio deepfake detection task.

$$\alpha_{\text{NER}} = \text{softmax} \left(\frac{\mathbf{Q}_{\text{NER}} \mathbf{K}_{\text{NER}}^{\top}}{\sqrt{D}} \right), \quad (5)$$

$$\mathbf{H}_{\text{NER-attended}} = \alpha_{\text{NER}} \mathbf{V}_{\text{NER}}. \quad (6)$$

D. Attention Fusion

The attention fusion mechanism merges $\mathbf{H}_{\text{ADD-attended}}$ and $\mathbf{H}_{\text{NER-attended}}$ through a gated attention strategy. Specifically, the fusion weight g is computed as

$$g = \sigma(W_g [\mathbf{H}_{\text{ADD-attended}}, \mathbf{H}_{\text{NER-attended}}]), \quad (7)$$

where σ denotes the Sigmoid function, and $[\cdot, \cdot]$ represents concatenation. Because of the Sigmoid activation, the fusion weight g is constrained to the range $[0, 1]$, where a value of 0

indicates full reliance on $\mathbf{H}_{\text{NER-attended}}$, a value of 1 indicates full reliance on $\mathbf{H}_{\text{ADD-attended}}$, and intermediate values represent a weighted combination of the two.

The fused features are then calculated as

$$\mathbf{H}_{\text{fused}} = g \cdot \mathbf{H}_{\text{ADD-attended}} + (1 - g) \cdot \mathbf{H}_{\text{NER-attended}}. \quad (8)$$

which are subsequently fed into the backend classifier.

E. Attention Transfer

The attention transfer mechanism aligns α_{ADD} with α_{NER} , enabling the teacher weights α_{NER} to guide the student weights α_{ADD} in learning semantic information. Specifically, this alignment is achieved by minimizing the KL divergence between the two distributions:

$$\mathcal{L}_{\text{KL}} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^T \alpha_{\text{NER},j} \log \frac{\alpha_{\text{NER},j}}{\alpha_{\text{ADD},j}}. \quad (9)$$

Due to its asymmetry, KL divergence is employed in a teacher–student manner: α_{NER} serves as the teacher and α_{ADD} as the student. Minimizing $KL(\alpha_{\text{NER}} \parallel \alpha_{\text{ADD}})$ ensures one-way semantic knowledge transfer from SpeechNER to PADD.

The total loss combines the primary binary cross-entropy (BCE) loss with the transfer loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}}. \quad (10)$$

Here, λ_{KL} is a hyperparameter that balances the two components. This new loss guides the update of α_{ADD} , which is then used to compute attention on \mathbf{H}_{ADD} , producing $\mathbf{H}_{\text{ADD-attended}}$, the input to the backend classifier.

F. Backend Classifier

The embeddings are passed to the frame-level backend classifier for final prediction. To capture long-range global context within each frame, we employ multiple Transformer encoders. Subsequently, we apply a BiLSTM to further model the sequence embeddings generated by the Transformer encoders. Finally, a fully connected layer is used to predict the boundary probability for each frame.

III. EXPERIMENTS AND RESULTS

A. Data Preparation

To rigorously validate the proposed method, we constructed the PartialSpoof-NER dataset, which incorporates named entity information. This dataset was derived from the PartialSpoof dataset by leveraging a multi-step process. Specifically, the audio data from the PartialSpoof dataset was first transcribed into text using the ASR model Whisper [21]. Subsequently, named entities were extracted from the transcriptions using Stanza³, a tool developed by the Stanford University research team. The resulting dataset, annotated with named entity information, was then curated as PartialSpoof-NER.

Table I presents the statistics of the datasets used in our experiments. As shown in Table I, the Train dataset consists

³<https://github.com/stanfordnlp/stanza>

TABLE I
THE STATISTICS OF DATASETS.

Name	Bona fide	Fake	All	Named Entities Count
Train	966	8789	9755	11572
Dev	124	1057	1181	1407
Eval	122	1126	1248	1526

TABLE II
COMPARISON OF EER (%) RESULTS AMONG DIFFERENT MODELS, INCLUDING OUR PROPOSED NE-PADD FRAMEWORK (REPORTED RESULTS REPRESENT NE-PADD-AF AS THE OPTIMAL IMPLEMENTATION).

Model	EER
Single reso. [18]	26.51
BAM [23]	18.21
TDL [24]	14.64
WBD [19]	11.59
NE-PADD	7.89

of 966 bona fide utterances and 8,789 fake utterances, with a total of 11,572 named entities. The Dev dataset contains 124 bona fide utterances and 1,057 fake utterances, with a total of 1,407 named entities. The Eval dataset includes 122 bona fide utterances and 1,126 fake utterances, encompassing 1,526 named entities in total. Each dataset contains, on average, 1 to 2 named entities per utterance.

B. Experiment Setup

The model is optimized using binary cross-entropy loss with the Adam optimizer over 100 epochs. Training is conducted with a mini-batch size of 16 and a learning rate of 10^{-4} , with a Noam scheduler [22] applied for learning rate adjustment, including 1,600 warm-up steps. During training, the equal error rate (EER) is periodically evaluated on the adaptation set to monitor model performance.

In NE-PADD, the input audio is represented with dimensions $L \times 1$, where L denotes the audio length. The parameters of the SpeechNER feature extraction module are frozen. In the PADD feature extraction module, the input is processed with Wav2Vec, and the output is transformed to a size of $T \times 768$, where T represents the number of time stamps. The first 1D convolutional layer is configured with $C(5, 2, 1)$, without bias, producing an output of $T \times 512$. This is followed by 12 residual blocks, each consisting of convolution and normalization layers with $C(1, 0, 1)$, maintaining the output size at $T \times 512$. A second 1D convolutional layer with the structure $C(1, 0, 1)$ further reduces the output size to $T \times 128$.

The backend classifier uses a Transformer encoder with $E(2, 4, 1024)$, which consists of 2 layers, 4 attention heads, and a feedforward network size of 1024, maintaining the output dimensions at $T \times 128$. Finally, a BiLSTM layer with 1 hidden layer of size 128 produces an output of $T \times 128$, followed by a ReLU activation function; ultimately, a 256-d fully-connected layer is used to predict the probability of each frame.

TABLE III
COMPARISON OF EXPERIMENTAL RESULTS AFTER APPLYING ATTENTION FUSION (AF) AND ATTENTION TRANSFER (AT) TO EACH BASELINE MODEL.

Model	EER
BAM-AT	14.84
BAM-AF	10.61
TDL-AT	8.87
TDL-AF	9.28
NE-PADD-AT	9.48
NE-PADD-AF	7.89

C. Baselines

To demonstrate the effectiveness of the proposed NE-PADD model, we compared it against four state-of-the-art PADD models: **1) Single reso.** [18] innovates in architecture by introducing SSL models and training strategies using segment-level labels for multi-resolution training in the PS spoofing scenario. **2) BAM** [23] utilizes boundary information as auxiliary attention for partially spoofed audio localization and develops a two-branch boundary module to exploit frame-level information for discriminative features. **3) TDL** [24] combines an embedding similarity module and temporal convolution operation to effectively capture both feature and positional information. **4) WBD** [19] proposes a frame-level boundary detection system using segment discontinuity and acoustic information.

D. Metrics

We use the Equal Error Rate (EER) to evaluate our model, a standard metric in binary classification tasks. Although our task involves frame-level partial spoof detection, each frame prediction can still be treated as a binary decision, making EER a relevant and informative measure. The EER corresponds to the point where the false positive rate equals the false negative rate, providing a balanced assessment of detection performance. A lower EER indicates higher overall accuracy in distinguishing between genuine and spoofed frames.

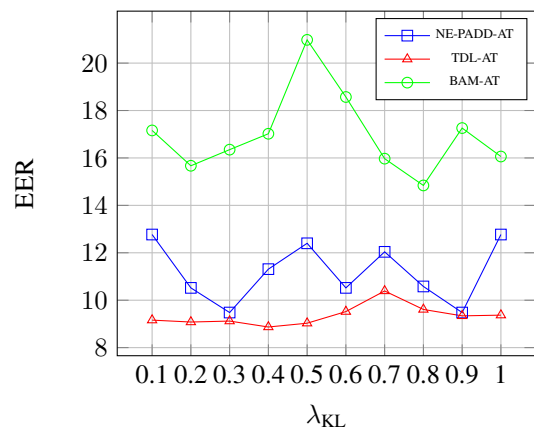


Fig. 2. EER Results for Different Values of λ_{KL}

E. Comparison with Existing Methods

As shown in Table II, our attention fusion method achieved an EER of just 7.89%, achieving the best performance on the PartialSpoof-NER dataset. It is worth clarifying that NE-PADD refers to our overall framework that leverages named entity semantic information to enhance partial audio deepfake detection. Within this framework, we explore two concrete implementation strategies: NE-PADD-AF and NE-PADD-AT. Both strategies follow the same core idea of exploiting speech-NER semantic weights, but differ in how the attention signals are integrated. Since NE-PADD-AF delivers the best empirical results, we report its performance as the representative implementation of NE-PADD when comparing with other baselines. By introducing these innovative attention fusion and attention transfer methods under the NE-PADD framework, our model achieves more precise capture of named entities' semantic weights, leading to superior outcomes in the PADD tasks.

F. Ablation study

We conducted ablation experiments to evaluate the effectiveness of integrating semantic information in enhancing the performance of PADD models. By applying our proposed attention fusion and attention transfer methods to various models, we enabled them to learn semantics directly from audio signals. As shown in Table III, experimental results reveal that the BAM model achieved significant EER reductions of 3.37% and 7.60% when incorporating the attention transfer and attention fusion methods, respectively, compared to the baseline model. Similarly, the TDL model exhibited EER decreases of 5.77% and 5.36% through the implementation of attention transfer and attention fusion methods. These findings demonstrate that introducing semantic information via attention mechanisms substantially improves detection capabilities in PADD models. The results underscore the critical role of semantic features in identifying subtle audio manipulations and validate the efficacy of integrating such features into the detection framework. This research highlights the importance of leveraging semantic understanding for combating increasingly sophisticated audio forgery techniques. For clarity, when we denote variants such as BAM-AT and BAM-AF, or TDL-AT and TDL-AF, they refer to the baseline BAM (or TDL) model augmented with our AT or AF mechanisms, respectively. This naming convention is consistent with the NE-PADD framework described earlier.

For the attention transfer method, as shown in Fig. 2, we conducted multiple experiments on three models by varying the hyperparameter λ_{KL} in steps of 0.1, ranging from 0.1 to 1. The results for attention transfer presented in Table III reflect the optimal hyperparameter values for each model.

G. Finer-grained resolution experiment

To investigate the impact of the number of spoofed segments in an audio file on the model's detection performance, we split the existing dataset by the number of spoofed segments (e.g., 1 forgery, 2 forgeries, 3 forgeries, up to 10 forgeries) and trained and tested the model on these subsets. As shown in Fig. 3, while the limited size of the dataset may prevent the results

from fully reflecting the model's generalization ability, we observed that as the number of spoofed segments increased, the model's accuracy in detecting forgeries improved, as evidenced by a lower EER. This indicates that the number of spoofed segments has a significant impact on model performance.

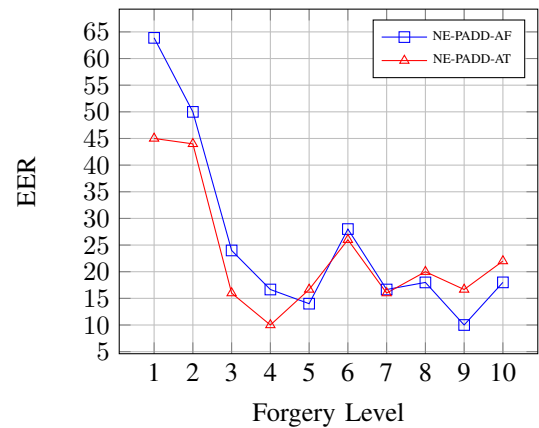


Fig. 3. EER results for the dataset with different levels of segmentation forgery.

IV. CONCLUSIONS

In this paper, we propose a novel method for partial deepfake audio detection and introduce PartialSpoof-NER, a new dataset. Our model uses two attention mechanisms to fuse semantic features, improving detection of subtle audio manipulations. It achieves state-of-the-art performance on PartialSpoof-NER. Detecting very short spoofed segments remains challenging and will be a focus of future work. Due to limited data, the model has not yet been applied to Chinese-English partial forgery; future work will expand data resources to validate multilingual generalization.

V. ACKNOWLEDGMENT

The research by Rui Liu was funded by the Young Scientists Fund (No. 62206136), the General Program (No. 62476146) of the National Natural Science Foundation of China, the Young Elite Scientists Sponsorship Program by CAST (2024QNRC001), the Outstanding Youth Project of Inner Mongolia Natural Science Foundation (2025JQ011), Key R&D and Achievement Transformation Program of Inner Mongolia Autonomous Region (2025YFHH0014) and the Central Government Fund for Promoting Local Scientific and Technological Development (2025ZY0143). The work of Huhong Xian was funded by the Research and Innovation Project for Graduate Students of Inner Mongolia University. The work by Berrak Sisman was supported by NSF CAREER award IIS-2338979. The work by Haizhou Li was supported by the Shenzhen Science and Technology Program (Shenzhen Key Laboratory, Grant No. ZDSYS20230626091302006), the Shenzhen Science and Technology Research Fund (Fundamental Research Key Project, Grant No. JCYJ20220818103001002), and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams, Grant No. 2023ZT10X044.

REFERENCES

- [1] W. Wu, Z. Lin, Y. Zhou, *et al.*, “Diffcss: Diverse and expressive conversational speech synthesis with diffusion models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.
- [2] W. Han, M. Kang, C. Kim, and E. Yang, “Stable-tts: Stable speaker-adaptive text-to-speech synthesis via prosody prompting,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.
- [3] M. Kang, D. Min, and S. J. Hwang, “Grad-stylespeech: Any-speaker adaptive text-to-speech synthesis with diffusion models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.
- [4] J. Li, W. Tu, and L. Xiao, “Freevc: Towards high-quality text-free one-shot voice conversion,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [5] C. H. Chan, K. Qian, Y. Zhang, and M. Hasegawa-Johnson, “Speechsplit2.0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6332–6336.
- [6] Y. Chen, D. Wu, T. Wu, and H. Lee, “Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5954–5958.
- [7] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, “Avqvc: One-shot voice conversion by vector quantization with applying contrastive learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4613–4617.
- [8] J. Zuo, S. Ji, M. Fang, *et al.*, “Enhancing expressive voice conversion with discrete pitch-conditioned flow matching model,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.
- [9] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 33, 2020, pp. 17 022–17 033.
- [10] A. Van Den Oord, S. Dieleman, H. Zen, *et al.*, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, vol. 12, p. 1, 2016.
- [11] J. Shen, R. Pang, R. J. Weiss, *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.
- [12] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, “End-to-end anti-spoofing with rawnet2,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6369–6373.
- [13] J.-w. Jung, H.-S. Heo, H. Tak, *et al.*, “Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2022, pp. 6367–6371.
- [14] J. Yi, Y. Bai, J. Tao, *et al.*, “Half-truth: A partially fake audio detection dataset,” *arXiv preprint arXiv:2104.03617*, 2021.
- [15] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino, and N. Evans, “An initial investigation for detecting partially spoofed audio,” *arXiv preprint arXiv:2104.02518*, 2021.
- [16] L. Zhang, X. Wang, E. Cooper, and J. Yamagishi, “Multi-task learning in utterance-level and segmental-level spoof detection,” *arXiv preprint arXiv:2107.14132*, 2021.
- [17] A. Baeovski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems (NIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 12 449–12 460.
- [18] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, “The partialspool database and countermeasures for the detection of short fake speech segments embedded in an utterance,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2022.
- [19] Z. Cai, W. Wang, and M. Li, “Waveform boundary detection for partially spoofed audio,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.
- [20] H. Yadav, S. Ghosh, Y. Yu, and R. R. Shah, “End-to-end named entity recognition from english speech,” *arXiv preprint arXiv:2005.11184*, 2020.
- [21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning (ICML)*, PMLR, 2023, pp. 28 492–28 518.
- [22] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” *Advances in neural information processing systems (NIPS)*, vol. 30, 2017.
- [23] J. Zhong, B. Li, and J. Yi, “Enhancing partially spoofed audio localization with boundary-aware attention mechanism,” *arXiv preprint arXiv:2407.21611*, 2024.
- [24] Y. Xie, H. Cheng, Y. Wang, and L. Ye, “An efficient temporary deepfake location approach based embeddings for partially spoofed audio detection,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 966–970.