Interspeech 2024
1-5 September 2024, Kos, Greece

HarmoNet: Partial DeepFake Detection Network based on Multi-scale HarmoF0 Feature Fusion

Liwei Liu[1], *Huihui Wei*[2], *Dongya Liu*[1], *Zhonghua Fu*[1,*]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University (NPU), Xi'an, China
[2]Shanghai Jiao Tong University, Shanghai, China

lwliu@mail.nwpu.edu.cn, huihuiwei@sjtu.edu.cn, mailfzh@nwpu.edu.cn

Abstract

Audio DeepFake detection (ADD) has become an increasingly challenging task recently, with the rise of various spoofing attacks utilizing artificially generated audio. The track 2 of ADD 2023 requires not only detecting DeepFake audio but also locating the manipulated regions. To tackle this unique challenge, we have proposed an innovative framework *HarmoNet* that leverages the Multi-scale harmonic F0 and Wav2Vec features with attention mechanism. This allows the model to effectively capture changes in each region of the utterance. Furthermore, we have introduced a new loss function named Partial Loss, which focuses more on the boundary between real and fake region. Additionally, we have designed a post-processor to refine the output of the model. Our framework achieved 70.61% in track 2 of ADD 2023, an improvement of 67.12% over baseline, and achieved the best performance. Moreover, *HarmoNet* also shows competitive performance on other DeepFake datasets.

Index Terms: Audio DeepFake, Self-supervised Learning, Harmonic F0, Partial Loss.

1. Introduction

As the core technologies of intelligent speech, automatic speech recognition (ASR) [1] and automatic speaker verification (ASV) [2] bring potential security problems and pose the risk of being attacked. The attacker uses a variety of technologies such as text-to-speech (TTS) [3] and voice conversion (VC) [4] to produce DeepFake audio, and thus gain access to a voice-controlled system. Therefore, detecting the authenticity of the audio has become an urgent and challenging task.

A series of challenges have played a critical role in this area, such as ASVspoof series [5, 6] and ADD 2022 [7]. Unlike the past challenges, Track 2 in the second Audio Deepfake Detection Challenge (ADD 2023) [8] is more focused on the partially manipulated region of audio. This new task makes fake audio more covert and significantly increases the difficulty of detection. It may lead to tampering with the meaning of speech, as shown in Figure 1. Meanwhile, the test set of ADD 2023 comprises unseen data, including real audio from different domains and synthetic audio clips created by models which are different from the ones used in the training set.

Most previous studies extract various features of speech and then send them to the classifier module to distinguish the differences between real and fake audio. There are many features have been introduced, such as Mel-frequency Cepstral Coefficients (MFCC) [9], Linear Prediction Cepstral Coefficients (LPCC) [10], Constant Q Cepstral Coefficients (CQCC) [11], Linear Frequency Cepstral Coefficients (LFCC) [12] and so on. The fundamental frequency (F0), as an important component of speech, has also been used in many studies to detect DeepFake audio. Patel [13] fused F0 contour and MFCC at score level. Pal [14] emphasized pitch variation features when detecting synthetic speech. Xue [15] used the 0-400Hz frequency band as the F0 subband frequency and fully studied the effect of the F0 sub-band. Fan [16] integrated the spatial reconstructed local attention Res2Net and F0 sub-band to obtain Multi-scale information. These studies indicated that F0 significantly contributes to discriminating DeepFake audio from the genuine.

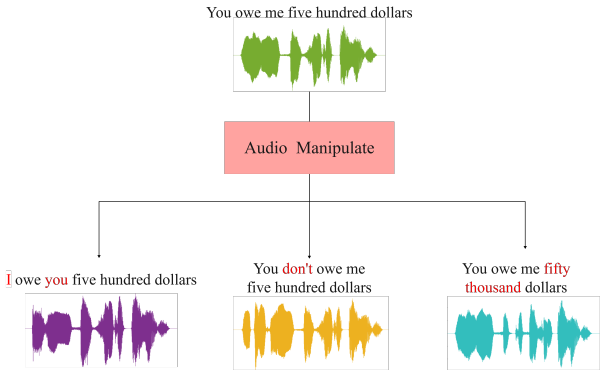

Figure 1: *Partially fake utterance changing a few words.*

Meanwhile, the development of self-supervised learning (SSL) enables models to learn richer and more latent representations. Some studies use self-supervised pre-trained models as the front-end to extract features [17, 18, 19], effectively demonstrating the usefulness of self-supervised learning for identifying fake audio. However, it is challenging to distinguish between partially fake audio and real audio by directly employing SSL models.

Motivated by both F0 and self-supervised learning, we propose HarmoNet for audio manipulation region localization in the ADD 2023 challenge Track 2. Our proposed model utilizes the latent representations extraction capability of SSL, along with the harmonic F0 (abbreviated as HarmoF0) characteristic of speech, to distinguish between real and fake audio.

The rest of the paper is organized as follows. In Section 2, we introduce the motivations behind using SSL and Harmonic information. In Section 3, we introduce the specific structure of HarmoNet, including HarmoF0 encoder, SSL encoder, feature fusion module, post processor and loss function. Next, we introduce the experiment, including datasets, baseline system, and metrics in Section 4. Then we analyze the results of the proposed model in Section 5. Finally, we get the conclusions in Section 6.

*Corresponding author.

10.21437/Interspeech.2024-1164

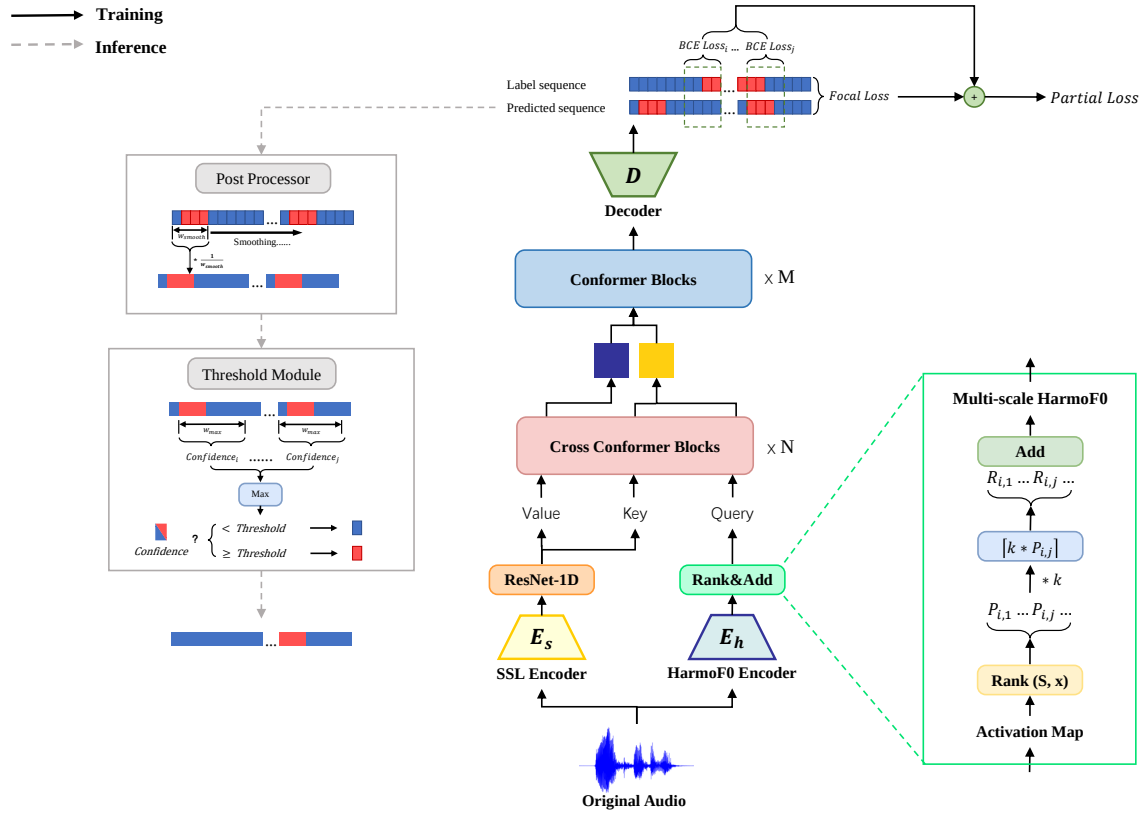

Figure 2: *Schematic diagram of HarmoNet.*

2. Motivation

Given the varying organ structures among individuals, each person has a unique fundamental frequency (F0) and harmonics. To make the generated sound more realistic, speech generation techniques usually simulate the F0 and harmonic of humans as realistically as possible. However, due to various factors such as individual differences, emotional states, language habits, and the accuracy of the neural network models, accurately replicating the intricate changes in F0 and harmonic remains challenging. Furthermore, the limited dataset of speech data used in TTS or VC systems affects the accuracy of F0 and harmonic simulation. Consequently, the F0 and harmonic can serve as a feature to assess the authenticity of speech signals. Therefore, based on F0 and harmonic related research, we designed Multi-scale HarmoF0 to obtain ordered and larger range harmonic information after obtaining HarmoF0 features. We believe that this feature can more effectively locate fake regions in audio.

Self-supervised learning (SSL) allows models to autonomously learn rich and meaningful generic speech feature representations from unlabeled data. Then fine-tune the pre-trained model with a small amount of supervised data, which can better assist downstream tasks such as ASR. Wav2Vec 2.0 [20] is a well-known self-supervised model that has achieved good results in several tasks, including Audio Deepfake Detection [18, 21]. Inspired by the success of Wav2Vec, we want to leverage the powerful feature extraction capacity of SSL, alongside the distinctive Multi-scale HarmoF0 characteristics, to effectively distinguish between real and fake audios.

Although SSL features have performed well in the previous DeepFake detection task of detecting real and fake binary classification, latent representations extracted from wav2vec may have low temporal resolution and cannot accurately locate the changes in partially fake audio. Therefore, we decide to use feature fusion. When employing attention mechanism for feature fusion, the HarmoF0 features act as a rough detection of speech activity and are able to reduce the effect of mute frames in the audio. Therefore, this fusion enables us to achieve a high level of accuracy in discerning between real and fake audio.

3. Proposed Approaches

In this section, we introduce our detection framework named *HarmoNet* for audio manipulation region location (RL). The structure mainly includes the HarmoF0 encoder, the SSL encoder, the feature fusion module, the sequence decoder, the proposal post-processor, and the loss function. The detailed structure of the HarmoNet framework is in Figure 2.

Firstly, an SSL pre-trained model is utilized as a SSL encoder to extract general representations of original speech signals, such as Wav2Vec 2.0. Meanwhile, a HarmoF0 encoder is used to capture the F0 and harmonic features. Then, some modified Conformer Blocks (Convolution-augmented Transformer) [22] are employed for feature fusion, enabling the model to efficiently distinguish between real and fake audio. The fused feature vectors are input into some normal Conformer Blocks for training. Finally, a bidirectional-LSTM (BiLSTM) is used to decode the context information and obtain the corresponding prediction sequence.

3.1. HarmoF0 Encoder

In this part, we utilize a dilated convolution network model similar to [23] to capture both the fundamental frequency (F0) and the harmonic characteristics of speech signals. This process can be seen as a way to highlight and emphasize phonetic characteristics of the original audio that are important for our task.

The original F0 model described in [23] only obtains F0 for each frame, which is specifically represented by the maximum predicted value of the activation map. However, what we need is the F0 for each frame and a series of harmonics (called Multi-scale HarmoF0). To extract Multi-scale HarmoF0, We have made some modifications to the output of the model. At first, we get the whole activation map output by the model and obtained the top *H* local maxima through ranking. Then we select the *H* frequency sub-bands according to the index of the local maxima, to get Multi-scale HarmoF0 features. Subsequently, the Multi-scale HarmoF0 are utilized as input features for the feature fusion module.

3.2. SSL Encoder

In terms of SSL encoder, we utilize the chinese-wav2vec2-large module[1], which pre-trained on 10,000 hours of Chinese data [24]. A detailed description of the Wav2Vec architecture can be found at [20]. After encoder, a residual network (ResNet-1D) layer is used to take the output of the Wav2Vec layer and transform it into a new SSL feature that had lower feature dimension.

3.3. Modified Conformer Block

For feature fusion, we modify the first feedforward module and multi-head self-attention module in the original Conformer Block, to fuse the SSL features with Multi-scale HarmoF0 features.

3.3.1. Dual Feed Forward Module

The dual feedforward module (DFM) has an additional data shunt compared to the original Conformer feedforward structure. In this module, the Multi-scale HarmoF0 feature is treated as Query, while the output of the ResNet-1D serves as both Key and Value. Since both the Key and Value originate from the same vector, a single path is used for training, while the other path is employed for training the Query.

3.3.2. Modified multi-head attention

After the DFM, we employ the cross attention mechanism [25] to fuse the SSL features with Multi-scale HarmoF0 features:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q (Query) correspondence of the Multi-scale HarmoF0 feature, K (Key) and V (Value) correspondence of the Wav2Vec feature.

3.4. Post Processor

We observe that if the model focus too much on the authenticity of each segment in the speech, the entire speech will be easily identified as fake due to the few frames that are incorrectly identified as fake. Therefore, we design a post-processor to improve the model accuracy using the smoothed probabilities.

$$p'_{ij} = \frac{1}{j - h_{\text{smooth}} + 1} \sum_{k=h_{\text{smooth}}}^{j} p_{ik} \tag{2}$$

where $h_{\text{smooth}} = \max\{1, j - w_{\text{smooth}} + 1\}$ is the index of the first frame within the smooth window, w_{smooth} is the window length, p_{ik} represents the probability of predicting the j-th frame as category i, where i is either real or fake. Take the average of the probabilities of the previous $w_{\max}$ frames for each label of the current frame. The confidence of each speech is then set to its highest post-smoothed probability, if the set threshold (empirical value) is exceeded, the predicted probability is used, if not, all frames are set to true.

3.5. Partial Loss Function

In order to address the problem of imbalanced data and make the model focus more on the boundary between real and fake region, we propose Partial Loss. This loss function is a weight combination of Focal Loss [26] and the Binary Cross Entropy (BCE) Loss in the boundary between real and fake regions:

$$\begin{aligned} PartialLoss &= w * FocalLoss \\ &+ (1 - w) * \sum_{f_i}^{M} \sum_{f_i - s}^{f_i + s} BCELoss \end{aligned} \tag{3}$$

where $w \in (0, 1)$, f_i represents the i-th boundary between real and fake regions in a sequence, M is the total number of boundaries, s represents that the BCE Loss is also computed for the s frames before and after the boundary. If there are no fake region in the current sequence, the loss function degenerates into Focal Loss. Generally, Focal Loss decreases the weight of the loss for easy samples in the overall loss, enabling the model to prioritize optimization of the hard samples.

4. Experiment

4.1. Dataset

Our training data is derived from the ADD 2023 track 2 (RL), focusing on locating the manipulated regions in partially fake audio, in which the original utterances are manipulated with real or generated audio. The training, dev, and test sets of RL are summarized in Table 1.

Table 1: *Numbers of the training, dev, and test datasets*

Name	Real	Fake	PartialFake	ALL
Train	26,554	1,185	25,354	53,093
Dev	8,914	430	8,480	17,824
Test	20,000	-	30,000	50,000

In order to improve our models' robustness for challenging conditions, we used data augmentation techniques during the training phase. We increase training data by 10% with noise from the MUSAN database [27]. Another 10% of data is augmented with music noise from MUSAN. Additionally, fragment operation is applied, with 10% of the data randomly selecting fragments for pitch shifting, and 10% of the data randomly selecting positions to insert fake fragments, with fragment lengths between 0.4s and 1.0s.

[1] https://github.com/TencentGameMate/chinese_speech_pretrain

Table 2: *ADD 2023 Track 2 Rankings. The results were performed by Final Score and Increase Rate*

Method	Score(%)↑	IncreaseRate(%)↑
Baseline	42.25	-
HarmoNet	**70.61**	**67.12**
C1[29]	67.13	58.89
C2[30]	62.49	47.90
C3[31]	62.02	46.79

Table 3: *The Performance of different feature combinations*

Feature	$A_{sen}(\%)\uparrow$	$F1_{score}(\%)\uparrow$	Score(%)↑
SSL	66.65	60.02	62.01
Multi-scale HarmoF0	72.18	55.44	60.46
LFCC + SSL	73.33	53.48	59.44
CQCC + SSL	**77.03**	62.07	66.56
MFCC + SSL	75.23	61.06	65.31
Multi-scale HarmoF0 + SSL	76.01	**68.31**	**70.61**

4.2. Baseline System and Evaluation metrics

We employ Final Score metrics to measure the effectiveness of anti-spoofing, as elaborated in [8]. The Final Score comprises the combined values of 30% Sentence Accuracy (A_{sen}) and 70% Segment F1-score ($F1_{score}$). A_{sen} is to measure the model's capacity to distinguish between real and fake audio, while $F1_{score}$ assesses the model's capability to identify fake region within DeepFake audio.

The baseline system is based on the LFCC-LCNN system [28], and the Final Score is 42.25%. To better evaluate the performance of the proposed model, we calculated the increase rate relative to the baseline system.

4.3. Experimental Setup

Our models are trained using the Adam optimizer for a total of 100 epochs. The batch size is set to 64. The initial learning rate is 0.005, it decays to the original 0.5 after every 20 epochs.

5. Results and Discussion

Firstly, we compare our HarmoNet with models in the challenge ADD 2023 Track 2. Our method achieves a score of 70.61%, a relative increase of 67.12% compared to the baseline system. We obtained the experimental results of other researchers from the official website of ADD 2023[2] [8] and our model achieved the best results. The ranking and Final Score of the models are shown in Table 2.

Secondly, to explore the role of various features in detecting DeepFake audio, we have set different feature combinations, including only SSL feature, only Multi-scale HarmoF0 feature, traditional features LFCC, CQCC, or MFCC fused with SSL feature, and Multi-scale HarmoF0 feature fused with SSL feature to train model. Noted that in HarmoNet that only uses single feature, there is no feature fusion module. The results are shown in Table 3. Compared with the LFCC, CQCC, and MFCC features, the fusion of Multi-scale HarmoF0 and SSL features achieves the highest Segment F1-score and Final Score, indicating the effectiveness of our framework. These features enable the model to pay more attention to F0 and harmonic-related information, which contain richer speech details. On the other hand, the Multi-scale HarmoF0 feature can be regarded as a basic voice activity detection component, which reduces the impact of silence frames on model training.

[2] http://addchallenge.cn/add2023

Table 4: *The Performance of top H in Multi-scale HarmoF0*

Feature	$A_{sen}(\%)\uparrow$	$F1_{score}(\%)\uparrow$	Score(%)↑
H = 1	72.90	61.04	64.59
H = 32	73.83	63.62	66.68
H = 64	75.19	64.43	67.66
H = 128	**76.01**	**68.31**	**70.61**

Table 5: *The performed with different loss functions*

Loss	$A_{sen}(\%)\uparrow$	$F1_{score}(\%)\uparrow$	Score(%)↑
BCE Loss	74.19	58.69	63.34
Focal Loss	75.76	60.98	64.42
Partial Loss	72.22	**68.38**	69.53
Partial Loss + Post processor	**76.01**	68.31	**70.61**

Our third experiment aims to verify the influence of *H* in Multi-scale HarmoF0. The results are shown in Table 4. It can be seen that as more harmonic information is extracted, the model's performance improves. When *H* = 1, Multi-scale HarmoF0 degenerates into F0, indicating that Multi-scale HarmoF0 can extract more information.

The last experiment aims to assess our model's performance with different loss functions. Specifically, we changed the loss function to BCE Loss, Focal Loss, proposed Partial Loss, and combined Partial Loss with post-processor to train the model. The results are shown in Table 5, which indicates that Partial Loss combines the advantages of BCE Loss and Focal Loss and achieves the best performance. The last line indicates that the post-processor module can alleviate the detection sensitivity of the model caused by a few error frames, and achieve a higher Final Score with a slight reduction in Segment F1-score.

Table 6: *The performance on the test set of Half-Truth dataset.*

Method	$A_{sen}(\%)$	$F1_{score}(\%)$	Score(%)
LCNN	89.59	82.66	80.07
HarmoNet	**90.60**	**90.91**	**90.82**

To assess our method's effectiveness, we also evaluated the HarmoNet model's performance on the Half-Truth dataset [32], which is a newly open partially fake dataset, the result is shown in Table 6. Our model demonstrates strong performance, indicating its effectiveness across different datasets.

6. Conclusion

In this paper, we propose *HarmoNet*, an audio DeepFake detection framework based on Wav2Vec and Multi-scale HarmoF0 features fusion, which takes advantage of the latent representations and the harmonic characteristics of speech signals. In addition, we design Partial Loss, to enhance the ability of the model to distinguish partially fake audio. Post-processor is set to reduce the impact of a few error frames. The experimental results show that HarmoNet can effectively identify and accurately locate the manipulated regions in partially fake audio. The model achieves an impressive score of 70.61% on the ADD2023 track2 test dataset, achieving the best performance, and its validity is verified on Half-Truth dataset.

7. References

[1] J. Li, "Recent advances in end-to-end automatic speech recognition," *ArXiv*, vol. abs/2111.01690, 2021.

[2] K. J. Devi and K. Thongam, "A survey of automatic speaker recognition system using artificial neural networks," *Journal of Advanced Research in Dynamical and Control Systems*, 2019.

[3] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *ArXiv*, vol. abs/2010.05646, 2020.

[4] J. Lian, C. Zhang, and D. Yu, "Robust disentangled variational speech representation learning for zero-shot voice conversion," in *ICASSP*. IEEE, 2022.

[5] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. W. D. Evans, T. H. Kinnunen, and K.-A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech*, 2019.

[6] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," 2021.

[7] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP*. IEEE, 2022.

[8] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, S. Nie, and H. Li, "Add 2023: the second audio deepfake detection challenge," *ArXiv*, vol. abs/2305.13774, 2023.

[9] K. A. Das, K. K. George, C. S. Kumar, S. Veni, and A. Panda, "Modified gammatone frequency cepstral coefficients to improve spoofing detection," in *ICACCI*, 2016.

[10] S. Jana, V. S. Yashwanth, K. N. Dheeraj, S. Balaji, K. P. Bharath, and M. R. Kumar, "Replay attack detection for speaker verification using different features level fusion system," *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pp. 1–5, 2021.

[11] M. Todisco, H. Delgado, and N. W. D. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Comput. Speech Lang.*, vol. 45, pp. 516–535, 2017.

[12] A. Chaiwongyen, K. Pinkeaw, W. Kongprawechnon, J. Karnjana, and M. Unoki, "Replay attack detection in automatic speaker verification based on resnewt18 with linear frequency cepstral coefficients," *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pp. 1–5, 2021.

[13] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (f0) and strength of excitation (soe) for spoofed speech detection," in *ICASSP*. IEEE, 2016.

[14] M. Pal, D. Paul, and G. Saha, "Synthetic speech detection using fundamental frequency variation and spectral features," *Comput. Speech Lang.*, vol. 48, 2018.

[15] J. Xue, C. Fan, Z. Lv, J. Tao, J. Yi, C. Zheng, Z. Wen, M. Yuan, and S. Shao, "Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features," *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022.

[16] C. Fan, J. Xue, J. Tao, J. Yi, C. Wang, C. Zheng, and Z. Lv, "Spatial reconstructed local attention res2net with f0 subband for fake speech detection," 2023.

[17] Z. Lv, S. Zhang, K. Tang, and P. Hu, "Fake audio detection based on unsupervised pretraining models," in *ICASSP*. IEEE, 2022.

[18] Z. Cai, W. Wang, and M. Li, "Waveform boundary detection for partially spoofed audio," *ArXiv*, vol. abs/2211.00226, 2022.

[19] C. Wang, J. Yi, J. Tao, C. Zhang, S. Zhang, and X. Chen, "Detection of cross-dataset fake audio based on prosodic and pronunciation features," *ArXiv*, vol. abs/2305.13700, 2023.

[20] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *ArXiv*, vol. abs/2006.11477, 2020.

[21] H. Wu, H.-C. Kuo, N. Zheng, K.-H. Hung, H. yi Lee, Y. Tsao, H.-M. Wang, and H. M. Meng, "Partially fake audio detection by self-attention-based fake span discovery," in *ICASSP*. IEEE, 2022.

[22] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," *ArXiv*, vol. abs/2005.08100, 2020.

[23] W. Wei, P. Li, Y. Yu, and W. Li, "Harmof0: Logarithmic scale dilated convolution for pitch estimation," vol. 2022-July. Proceedings - IEEE International Conference on Multimedia and Expo, 2022.

[24] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, D. Wu, and Z. Peng, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," 10 2021.

[25] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[26] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 318–327, 2017.

[27] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *ArXiv*, vol. abs/1510.08484, 2015.

[28] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," in *Interspeech*, 2020.

[29] Z. Cai, W. Wang, Y. Wang, and M. Li, "The dku-dukeece system for the manipulation region location task of add 2023," *ArXiv*, vol. abs/2308.10281, 2023.

[30] J. Liu, Z. Su, H. Huang, C. Wan, Q. Wang, J. Hong, B. Tang, and F. Zhu, "Transsionadd: A multi-frame reinforcement based sequence tagging model for audio deepfake detection," in *DADA@IJCAI*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259261939

[31] K. Li, X.-M. Zeng, J.-T. Zhang, and Y. Song, "Convolutional recurrent neural network and multitask learning for manipulation region location," in *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, vol. 750, 2023.

[32] J. Yi, Y. Bai, J. Tao, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-truth: A partially fake audio detection dataset," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233181848