INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features

Chenglong Wang[1,2], *Jiangyan Yi*[2,3], *Jianhua Tao*[3,4], *Chu Yuan Zhang*[2,3], *Shuai Zhang*[4], *Xun Chen*[1]

[1]University of Science and Technology of China, Hefei, China
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijng, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[4]Department of Automation, Tsinghua University

chenglong.wang@nlpr.ia.ac.cn

Abstract

Existing fake audio detection systems perform well in in-domain testing, but still face many challenges in out-of-domain testing. This is due to the mismatch between the training and test data, as well as the poor generalizability of features extracted from limited views. To address this, we propose multi-view features for fake audio detection, which aim to capture more generalized features from prosodic, pronunciation, and wav2vec dimensions. Specifically, the phoneme duration features are extracted from a pre-trained model based on a large amount of speech data. For the pronunciation features, a Conformer-based phoneme recognition model is first trained, keeping the acoustic encoder part as a deeply embedded feature extractor. Furthermore, the prosodic and pronunciation features are fused with wav2vec features based on an attention mechanism to improve the generalization of fake audio detection models. Results show that the proposed approach achieves significant performance gains in several cross-dataset experiments.

Index Terms: fake audio detection, ASVspoof, prosodic feature, pronunciation feature, cross-dataset

1. Introduction

Currently, various types of front-end features are used for detecting fake audio. These include short-time spectral features, raw audio, fundamental frequency features, and self-supervised features [1, 2, 3, 4, 5]. Todisco[6] demonstrated the superiority of Constant Q Cepstral Coefficients (CQCC) over Mel Frequency Cepstral Coefficients (MFCC) using the constant Q transform. Sahidullah[7] proposed Linear Frequency Cepstrum Coefficients (LFCC) by replacing Mel scale filters with linear filters, focusing on high-frequency band features. Das[8] improved CQCC features with Extended CQCC (eCQCC) and CQSPIC features. These methods achieved promising results in ASVspoof2019 logical access scenarios with EERs below 1

For prosodic features, fundamental frequency has been investigated [11, 12, 13, 14]. Prosodic features capture speech style and intonation from longer speech segments, such as phonemes and syllables [15]. Patel[16] improved fake audio detection by fusing F0 contour and 36-D MFCC at the score level. Pal[17] proposed fundamental frequency variation features to capture prosodic differences between real and fake audio. Xue[18] introduced F0 subband, utilizing F0 information for frequency band division. Another popular feature extractor is wav2vec, based on self-supervised learning [19, 20, 21]. Wav2vec features trained with unlabeled data have achieved top rankings in competitions [19, 21].

Detecting fake audio in practical applications can be challenging due to the limited generalization of existing systems to unknown types of spoofing attacks. This limitation arises from the lack of generalization capability in features extracted from a single dimension. For example, short-time spectral features, extracted from short frames of 20-30 ms, are susceptible to channel effects. Moreover, current methods for extracting prosodic features only focus on F0 features and overlook crucial phoneme duration features. It is worth noting that the duration of the same phoneme can vary significantly in different real audio contexts, while it tends to be more uniform in fake audio. Additionally, the self-supervised features generated by wav2vec may contain speech information that is difficult to identify.

To address these challenges, we propose the use of multi-view features to enhance the detection of fake audio, thereby improving generalization across datasets. Our approach incorporates features from three dimensions: prosodic, pronunciation, and wav2vec. We introduce phoneme duration extractors and pronunciation feature extractors to achieve this. To obtain phoneme-like duration features, we encode the speech using the pre-training model HuBERT[22] without the transcript. For pronunciation features, we train a Conformer-based[23] phoneme recognition model and utilize the acoustic encoder part as a deeply embedded feature extractor. To further enhance performance, we fuse prosodic and pronunciation features with discrete clustering-based wav2vec features using an attention mechanism. Our experimental results demonstrate that these auxiliary features improve the detection performance both within and outside the domain. The main contributions of this study can be summarized as follows:

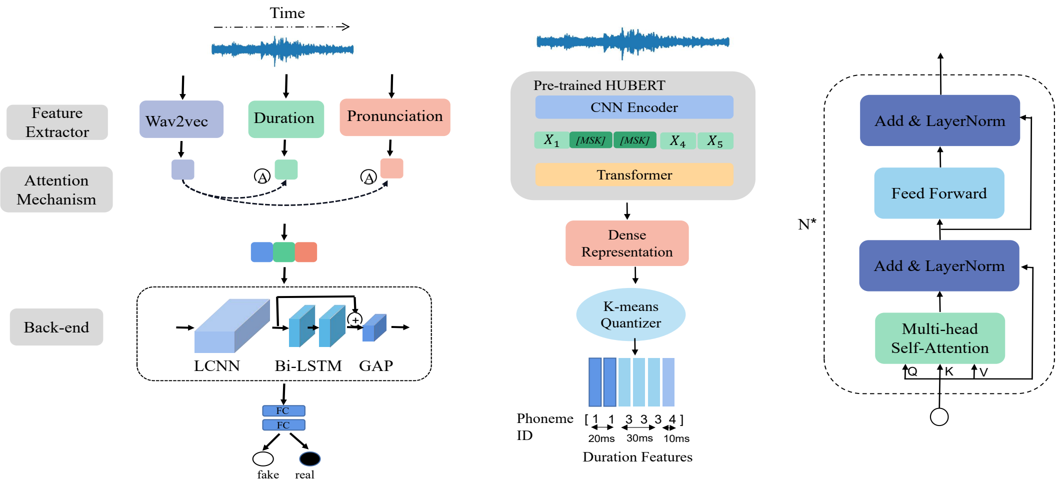
- We propose pronunciation features and phoneme duration features for fake audio detection for the first time.
- We use the attention mechanism approach to effectively fuse the prosodic features and pronunciation features with wav2vec features.

The rest of this paper is organized as follows: Section 2 illustrates our method. Experiments, results and discussions are reported in Section 3 and 4, respectively. Finally, we conclude the paper in Section 5.

2. Our Method

Our model comprises three modules: feature extractor, attention module, and back-end. The feature extractor extracts three types of features: wav2vec, phoneme duration, and pronunciation. We consider phoneme duration as a reflection of prosodic information, hence referring to it as ”prosodic feature.” The attention module fuses prosodic and pronunciation features with varying weights into the wav2vec features. The back-end module learns a deep speech representation. Figure 1 (a) illustrates

10.21437/Interspeech.2023-1254



(a) Overall Framework

(b) Duration Feature Extractor Module

(c) Attention Mechanism Module

Figure 1: *(a) Overall framework of proposed method. The system consists of the feature extraction module, LLGF block, and the attention mechanism module. Ⓐ denotes the attention mechanism. (b) Duration Encoder Module. The phoneme ID "134" in the figure refers to three different phonemes respectively. (c) The attention mechanism module.*

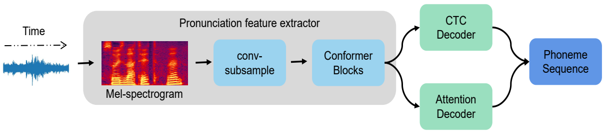


Figure 2: *Pronunciation Feature Extractor Module*

the architecture. Further details are provided in subsequent sections.

2.1. Features

Duration: Since existing publicly available fake audio datasets such as ASVspoof, ADD2022, etc., do not provide audio-corresponding transcript, we cannot extract the duration information of the fake audio by forcing the alignment. Inspired by [24], we encode the speech with the pre-training model HuBERT, and the resulting encoding vector is an encoding similar to speech phonemes. As shown in Figure 1 (b), the first step is to encode the original audio into an encoding vector with a HuBERT model which is pre-trained on the LibriSpeech corpus. We choose k-means as the quantization operation to transform the output of the encoder from continuous to discrete values. Formally, $D_i = Q(C_i)$, where Q is the quantization function k-means, C_i is a sequence of vectors, $D_i = [d_1, d_2, \ldots, d_T]$ such that $d_i \in \{1, 2, \ldots, K\}$ and K is the size of the phoneme vocabulary, we set $K = 100$. We refer to the final obtained D_i(e.g.,$[1, 1, 3, 3, 3, 4]$) as the phoneme duration vector.

Pronunciation: The Conformer model is widely used in speech recognition, so we adopted the pronunciation feature extractors based on the Conformer structure, as shown in Figure 2. First, we extract 80-dimensional log mel spectrograms from the raw audio. Then we use convolution downsampling in the time scale. The downsampled log mel spectrograms is then fed into the Conformer module, which follows the configuration in [23]. Then the predicted phoneme sequences are obtained through the CTC decoder and the attention decoder. The CTC decoder has a fully-connected layer. The attention decoder is location sensitive and has a decoder LSTM layer with a hidden size of 320. The training loss is a linear combination of the CTC and attention losses:

$$\mathcal{L} = \alpha\mathcal{L}_{CTC} + (1-\alpha)\mathcal{L}_{ATT} \tag{1}$$

Where $\mathcal{L}_{CTC}$ and $\mathcal{L}_{ATT}$ denote the loss of CTC and attention. $\alpha \in [0, 1]$ is a hyper-parameter, and we set it to be 0.5. After training, we keep the acoustic encoder part as a deeply embedded feature extractor. The pronunciation features encoded from raw audio are regarded as the pronunciation representations of the speech.

Wav2vec: Wav2vec 2.0 is a self-supervised speech representation learning method that can learn representations directly from raw audio signals without annotations. Its innovation lies in the use of the Transformer architecture, which can capture long-range dependencies. The model is trained using a masked contrastive predictive coding (CPC) objective to predict speech signals. The use of large-scale datasets and training processes improves the quality of the representations.

2.2. Fusion Strategy for Features

We use Transformer [25] to fuse the wav2vec feature with the other two features, respectively. We only use the encoder part of the Transformer. It is based on multi head attention mechanism. Multi head attention operates multiple self attention operations in parallel. The formula is as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

where d_k is the key dimension. Here, we use the embedding of wav2vec as keys and values, and the embedding of duration and pronunciation as queries, respectively. The multi-head attention mechanism obtains h different representations of (Q, K, V), computes scaled dot-product attention for each representation,

Table 1: *Statistics of experimental datasets.*

Datasets	#Real	#Fake	Others
ASVspoof2019(IN)	7,355	63,882	TTS and VC, English
ASVspoof2015(A)	9,404	184,000	TTS and VC, English
VCC2020(B)	2,660	6,120	VC, multilingual
In-the-Wild(C)	18,863	11,816	realistic, English
ADD2022(D)	30,000	70,000	partial fake, Chinese

concatenates the results, and projects the concatenation through a feed-forward layer. It can be defined as:

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^O \tag{3}$$

$$head_i = Attention(QW_i{}^Q, KW_i{}^K, VW_i{}^V) \tag{4}$$

2.3. Back-end Architecture

Regarding the back-end architecture, we follow the conclusion in [20] that a deep back-end is necessary when the front-end pre-training features are fixed. Our back-end architecture consists of a light convolutional neural network (LCNN) followed by two bi-directional recurrent layers with long short-term memory (LSTM) units, a global average pooling layer, and a fully connected (FC) output layer. We adopt the same configuration as presented in [20].

3. Experiments

3.1. Dataset

We employ five fake audio datasets. All of the models were trained on the ASVspoof2019[9] LA training sets. The ASVspoof2015[26] is the most similar to the ASVspoof2019 LA for their audios are collected from the same datasets or conversion algorithms. The VCC2020[27] dataset is multilingual. The In-the-Wild[28] dataset is collected from the real world. The ADD2022[29] track2 dataset is Chinese and it is partially fake.

3.2. Experimental Setup

The Wav2vec XLSR model was obtained from the Fairseq project[1]. The model was pre-trained on a training set that includes Multilingual LibriSpeech, CommonVoice, and BABEL, which cover 8, 36, and 17 languages, respectively. This extensive training corpus allows the Wav2vec XLSR model to learn robust and diverse speech representations across a wide range of languages and accents. In order to train the Conformer model described in sections 2.1, we utilized the LibriSpeech dataset, which consists of 960 hours of audio recordings [30]. The corresponding text transcripts were used in conjunction with the LibriSpeech lexicon [2] to obtain phoneme sequences for training the model.

The Wav2vec XLSR model has a dimension of 1024. In order to reduce the computational complexity, we apply a fully connected layer to reduce the dimensionality of the input features to 128. To form batches, we fix the length of each sample to 500 frames by truncating or concatenating. Thus, the shapes of the resulting wav2vec, duration and pronunciation features are 500×128, 500×1 and 500×144 , respectively. The audio sampling rate is 16k. For comparison, the baseline uses LFCC extracted with a frame length of 20 ms, a frame shift of 10 ms, and a 512-point fast Fourier transform (FFT). Each LFCC frame vector has a dimension of 60, including static, delta, and deltadelta components. In addition, 500 frames of the input is also needed at inference time.

For feature fusion, we use two methods: concatenation and the attention mechanism. For the former, we have directly concatenated the wav2vec features and the other two features from the temporal dimension, resulting in feature shapes of $500 \times (128+1+144)$. For the latter, we first perform the attention transformation of wav2vec and the other two features separately, and then concatenate the obtained vectors to get the final representation. For the attention mechanism, we use 6 blocks and 8 heads.

To train the model, we use the Adam optimizer with a learning rate of 5×10^{-5}. The batch size is 32. The model is trained for 200 epochs. The EER [31] is used as the evaluation metric.

4. Results and Discussion

4.1. Baseline

Table 2 displays the results of individual features on different datasets. It's important to note that our models were trained solely on the ASVspoof2019 LA dataset and subsequently tested on various datasets. Testing across datasets poses a significant challenge. For example, LFCC achieved an EER of 4.86% on the ASVspoof2019 LA set but exhibited decreased performance on the other four datasets. In contrast, Wav2vec demonstrated better generalization with an EER of 6.59% on the ASVspoof2015 LA test set but performed poorly on the VCC2020, in_the_wild, and ADD2022 datasets. The variations in fake speech generation and recording environments contribute to these observations. Our findings underscore the motivation of our paper, which aims to enhance the generalization of detection models across datasets.

The second observation is that individual prosodic or pronunciation features exhibited poor performance on in-domain and out-of-domain tests. This could be attributed to the fact that prosodic features are one-dimensional, while wav2vec features are 128-dimensional and LFCC features are 60-dimensional. Due to their limited dimensionality, prosodic features contain less information compared to short-time spectral features and wav2vec features, which leads to the loss of acoustically relevant information in speech. In addition, the poor performance of pronunciation features could be due to the small size of the training data for the pronunciation extractor, which only contained 960 hours of data, compared to the 436k hours of training data for wav2vec.

4.2. Proposed Method Results

The last few rows of Table 2 show the results of the proposed approach across datasets. We can draw the following conclusions: first, for both in-domain and out-of-domain tests, the performance of combining the prosodic features and pronunciation features with the wav2vec features is better than that of using the wav2vec features alone. This shows that using the prosodic and pronunciation features as auxiliary features has a positive impact. Specifically, concatenating the two features with the wav2vec features yielded the most noticeable performance improvement, with the EER decreasing from 6.59% to

[1] https://github.com/pytorch/fairseq/tree/main/examples/wav2vec/xlsr
[2] http://www.openslr.org/11/

Table 2: *EER(%) of our proposed different systems in in-domain and out-of-domain testing, where 'IN' denotes 'ASVspoof2019 LA' and 'A', 'B', 'C', 'D' denotes 'ASVspoof2015', 'VCC2020', 'in_the_wild' and 'ADD2020 track2', respectively. O_1 denotes concatenating and O_2 denotes the attention mechanism. When utilizing a single feature as input, neither concatenation nor attention fusion is employed. Results are the average obtained from three runs of each experiment with different random seeds.*

Feature	**IN**		**A**		**B**		**C**		**D**	
	O_1	O_2	O_1	O_2	O_1	O_2	O_1	O_2	O_1	O_2
Pron	11.82	-	25.31	-	40.83	-	58.53	-	45.38	-
Duration	20.82	-	42.18	-	48.31	-	66.48	-	48.51	-
Duration + Pron	9.72	-	32.74	-	37.47	-	59.43	-	43.26	-
LFCC	4.86	-	28.15	-	35.62	-	62.53	-	43.62	-
Pron + LFCC	3.82	3.63	25.89	24.07	31.58	29.64	57.24	56.77	37.86	36.22
Duration + LFCC	4.35	4.04	28.05	25.42	32.93	32.41	60.08	58.91	41.24	39.79
Duration + Pron +LFCC	3.51	**3.18**	25.43	**21.86**	30.84	**28.36**	55.49	**54.22**	35.94	**34.82**
Wav2vec	3.16	-	6.59	-	19.33	-	43.81	-	35.19	-
Pron + Wav2vec	2.83	1.97	4.85	3.28	17.28	15.31	40.79	39.36	32.18	31.52
Duration + Wav2vec	3.08	2.44	5.33	4.25	18.05	17.54	42.15	41.37	34.27	32.68
Duration + Pron + Wav2vec	2.35	**1.58**	3.96	**3.08**	16.45	**14.76**	38.57	**36.84**	30.77	**29.53**

Table 3: *Compare with the system of cross-dataset testing recently proposed. 'IN' denotes 'ASVspoof2019 LA', 'A' denotes 'ASVspoof2015', and 'B' denotes 'VCC2020'.*

Methods	**IN**	**A**	**B**
Vanilla [32]	2.29	26.30	41.66
AUG [32]	2.92	16.25	30.51
MT-AUG [32]	3.41	22.10	28.85
ADV-AUG [32]	3.23	14.38	27.07
Duration + Pron +LFCC (ours)	3.18	21.86	28.36
Duration + Pron + Wav2vec (ours)	**1.58**	**3.08**	**14.76**

3.96% on the ASVspoof2015 test set. This is due to the addition of wav2vec information from the prosodic and pronunciation dimensions. Second, the fusion method using the attention mechanism has better performance than the direct concatenating method. For example, when the front-end input is "duration + pron+ wav2vec", the performance of the attention mechanism approach is 1.69% better than the concatenate operation on the VCC2020 test set. This is because the attention mechanism allocates the weights of prosodic features and pronunciation features, which makes the model pay more attention to features that can distinguish real and fake audio, thus improving the generalization of the model. Third, in the cross-dataset test, the proposed method improves ASVspoof2015 more than the other three datasets. For example, when the front-end input is "duration + pron+ wav2vec", the relative improvement of ASVspoof2015 with WAV2VEC features alone is 53.26%, compared with 23.64%, 15.90%, and 15.76% for VCC2020, in_the_wild, and ADD2022, respectively. This is because the dataset composition of ASVspoof2015 is similar to that of ASVspoof2019. Moreover, for LFCC, similar conclusions to wav2vec can be drawn. First, for both in-domain and out-of-domain tests, the performance of combining the prosodic features and pronunciation features with the LFCC features is better than that of using the LFCC features alone. Second, the fusion method using the attention mechanism has better performance than direct concatenating method. Third, in the cross-dataset test, the proposed method improves ASVspoof2015 more than the other three datasets.

Table 3 compares our proposed methods with recently proposed cross-dataset testing systems [32]. When utilizing "duration + pron + wav2vec" as the front-end feature, our system significantly outperforms other systems in both in-set and out-of-set performance. Although the performance improvement is not as pronounced as with "duration + pron + wav2vec," using "duration + pron + LFCC" still achieves competitive results in both in-set and out-of-set scenarios. Notably, we observed a substantial enhancement in cross-dataset testing when solely relying on wav2vec features. Since wav2vec features are exclusively trained on real speech data without exposure to fake audio, they are theoretically well-suited for generalizing to all types of fake audio. However, experimental findings indicate that the generalization of wav2vec features is still influenced by the match between the test and training sets. Our experiments demonstrate that fusing prosodic and pronunciation features with wav2vec features can further enhance the generalization of cross-dataset detection.

5. Conclusion

In this paper, we observe a significant performance degradation of existing fake audio detection systems in cross-dataset testing. This paper proposes multi-view features for fake audio detection, which attempts to capture more generalized features from the view of prosodic features, pronunciation features and wav2vec features. The results show that the prosodic and pronunciation features can be used as auxiliary features to improve the detection performance in and out of the domain. The fusion of the prosodic features and pronunciation features with wav2vec features is more effective by using the attention mechanism. In the future, we will explore different fusion strategies.

6. Acknowledgments

This work is supported by the National Key Research and Development Plan of China (No.2020AAA0140003), the National Natural Science Foundation of China (NSFC) (No.61831022, No.U21B2010, No.62101553, No.61971419, No.62006223, No.62276259, No.62201572, No. 62206278), Beijing Municipal Science&Technology Commission,Administrative Commission of Zhongguancun Science Park No.Z211100004821013, Open Research Projects of Zhejiang Lab (NO. 2021KH0AB06).

7. References

[1] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features." in *Interspeech*, 2017, pp. 27–31.

[2] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the asvspoof 2017 challenge." in *Interspeech*, 2017, pp. 7–11.

[3] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "Stc anti-spoofing systems for the asvspoof 2015 challenge," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5475–5479.

[4] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," Tech. Rep., 2016.

[5] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on asvspoof 2019," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1018–1025.

[6] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[7] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech 2015*, 2015, pp. 2087–2091.

[8] R. K. Das, J. Yang, and H. Li, "Assessing the scope of generalized countermeasures for anti-spoofing," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6589–6593.

[9] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

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

[11] X. Xiao, X. Tian, S. Du, H. Xu, E. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: the ntu approach for asvspoof 2015 challenge." in *Interspeech*, 2015, pp. 2052–2056.

[12] L. Attorresi, D. Salvi, C. Borrelli, P. Bestagini, and S. Tubaro, "Combining automatic speaker verification and prosody analysis for synthetic speech detection," *arXiv preprint arXiv:2210.17222*, 2022.

[13] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: from the perspective of asvspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, p. e2, 2020.

[14] H. A. Patil, M. R. Kamble, T. B. Patel, and M. H. Soni, "Novel variable length teager energy separation based instantaneous frequency features for replay detection." in *INTERSPEECH*, 2017, pp. 12–16.

[15] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

[16] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (f 0) and strength of excitation (soe) for spoofed speech detection," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5105–5109.

[17] M. Pal, D. Paul, and G. Saha, "Synthetic speech detection using fundamental frequency variation and spectral features," *Computer Speech & Language*, vol. 48, pp. 31–50, 2018.

[18] 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," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022, pp. 19–26.

[19] Z. Lv, S. Zhang, K. Tang, and P. Hu, "Fake audio detection based on unsupervised pretraining models," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9231–9235.

[20] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," *arXiv preprint arXiv:2111.07725*, 2021.

[21] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9241–9245.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[23] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[24] E. Kharitonov, J. Copet, K. Lakhotia, T. A. Nguyen, P. Tomasello, A. Lee, A. Elkahky, W.-N. Hsu, A. Mohamed, E. Dupoux *et al.*, "textless-lib: a library for textless spoken language processing," *arXiv preprint arXiv:2202.07359*, 2022.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth annual conference of the international speech communication association*, 2015.

[27] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z.-H. Ling, and T. Toda, "Voice conversion challenge 2020—intra-lingual semi-parallel and cross-lingual voice conversion," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98.

[28] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *arXiv preprint arXiv:2203.16263*, 2022.

[29] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[31] J.-M. Cheng and H.-C. Wang, "A method of estimating the equal error rate for automatic speaker verification," in *2004 International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 285–288.

[32] Y. Zhang, G. Zhu, F. Jiang, and Z. Duan, "An empirical study on channel effects for synthetic voice spoofing countermeasure systems," *Proc. Interspeech 2021*, pp. 4309–4313, 2021.