Voice Spoofing Detection via Speech Rule Generation Using wav2vec 2.0-Based Attention

Qian-Bei Hong
Department of Electrical Engineering, Southern Taiwan University of Science and Technology
qbhong75@gmail.com

Yu-Chen Gao
Department of Electrical Engineering, Southern Taiwan University of Science and Technology
4B12C100@office.stust.edu.tw

Yu-Ying Xiao
Department of Electrical Engineering, Southern Taiwan University of Science and Technology
4B127052@office.stust.edu.tw

Yeou-Jiunn Chen
Department of Electrical Engineering, Southern Taiwan University of Science and Technology
chenyj@stust.edu.tw

Kun-Yi Huang
Department of Computer Science and Information Engineering, Southern Taiwan University of Science and Technology
iamkyh77@stust.edu.tw

Abstract

Recent advancements in AI-based voice cloning have led to increasingly convincing synthetic speech, posing significant threats to speaker verification systems. In this paper, we propose a novel voice spoofing detection method that integrates acoustic feature variations with attention mechanisms derived from wav2vec 2.0 representations. Unlike prior approaches that directly utilize wav2vec 2.0 features as model inputs, the proposed method leverages wav2vec 2.0 features to construct speech rules characteristic of bona-fide speech. Experimental results indicate that the proposed RULE-AASIST-L system significantly outperforms the baseline systems on the ASVspoof 2019 LA evaluation set, achieving a 24.6% relative reduction in equal error rate (EER) and an 10.8% reduction in minimum tandem detection cost function (min t-DCF). Ablation studies further confirm the importance of incorporating speech rules and selecting appropriate hidden layer representations. These findings highlight the potential of using self-supervised representations to guide rule-based modeling for robust spoofing detection.

Keywords: Voice spoofing detection, Speech rule generation, wav2vec 2.0

1 Introduction

Telecom fraud has become a critically important issue today, particularly the method of using AI to synthesize the voices of victims' family members to impersonate them and commit financial fraud. This has emerged as a new tactic employed by scam groups. Voice spoofing can be primarily divided into two categories: Physical Access (PA) attacks and Logical Access (LA) attacks. According to past research in relevant literature, the difficulty of signal detection in LA attacks is typically greater than that in PA attacks. This is primarily because voice conversion and text-to-speech technologies can more accurately mimic the target speaker's voice, rather than merely reproducing recorded playback quality.

To address the growing threat of voice spoofing attacks, many studies have adopted deep neural network (DNN)-based models to classify speech as either genuine or spoofed (Y. Zhang et al., 2021; J. Zhou et al., 2022; A. Gomez-Alanis et al., 2019). However, these approaches typically treat spoofing detection as a binary classification problem that focuses solely on surface-level acoustic differences,

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 108–115

without accounting for the complexity and diversity of feature variations introduced by different spoofing methods. To address this limitation, (J. Boyd et al., 2023) proposed using a multi-class classification framework that distinguishes between genuine, voice conversion, speech synthesis, and replay categories. This enables the model to learn more discriminative features for identifying various types of spoofing attacks targeting genuine speech. However, most existing research on voice spoofing detection focuses on feature analysis from a single audio perspective.

Self-supervised learning (SSL) has emerged as a powerful alternative for extracting high-dimensional representations of speech signals (A. Baevski et al, 2020; W.-N. Hsu et al., 2021; S. Chen et al., 2022). These models typically rely on convolutional neural network (CNN)-based feature encoders, where CNN kernels perform nonlinear transformations on short segments of audio. A key advantage of self-supervised learning lies in its ability to learn from large-scale unlabeled data, enabling pre-trained models to capture a wide range of speech variability. Compared to conventional frequency-domain methods, these learned representations often yield more robust and informative features. Recently, SSL models such as wav2vec 2.0 (A. Baevski et al., 2020) have gained significant attention in various speech-related tasks. Originally developed for automatic speech recognition (ASR) (A. Bawitlung et al., 2025), these models have also demonstrated strong performance in speaker verification (Z. Fan et al., 2021) and speech emotion recognition (B. Nasersharif and M. Namvarpour, 2024). Recently, several studies have investigated the application of wav2vec 2.0 for spoofing detection tasks (H. Tak et al., 2022), taking advantage of its rich contextualized speech representations to improve feature modeling and detection accuracy.

This work proposes a novel framework that integrates conventional acoustic feature analysis with the sequential representation patterns derived from wav2vec 2.0. By exploring the interactions between acoustic features and the sequential rule of wav2vec 2.0 representations, the proposed approach enables voice spoofing detection not only from the inherent characteristics of speech but also through identifying inconsistencies in the sequence patterns of wav2vec 2.0 representations correlated with spoofed audio. This joint analysis enhances detection performance by uncovering unnatural patterns indicative of spoofed speech.

This paper addresses the fraudulent methods arising from current AI voice cloning technologies by proposing a detection method that combines the correlation between acoustic features and wav2vec 2.0-based attention mechanisms. This approach simultaneously considers the interaction between variations in acoustic features, and the rules of speech representations, aiming to enhance the accuracy of distinguishing between synthetic and genuine voices.

2 Related Work

2.1 AASIST

The AASIST network is composed of four main components: an encoder module, graph modules, a max graph operation (MGO) module, and an output module, as shown in the upper part of Figure 1. The encoder, based on RawGAT-ST (H. Tak et al., 2021), extracts high-level feature representations F directly from the raw audio waveform. Two parallel graph modules are employed to model the spectral and temporal characteristics of F, respectively, producing graph-structured features in both domains. These outputs are then fused to construct a heterogeneous graph, which is further processed by the MGO module.

The MGO module consists of two parallel upper and lower branches, each comprising two heterogeneous attention mechanisms and two stacked nodes that store time-frequency heterogeneous information. The final representation is obtained by applying an element-wise max operation to the outputs of the two branches. This representation is used to discriminate between bona-fide and spoofed speech.

2.2 wav2vec 2.0 Representations

wav2vec 2.0 leverages self-supervised learning to derive informative and high-level speech representations directly from raw audio input. Its architecture consists of two primary components: a convolutional feature extractor and a Transformer-based contextual module. The convolutional encoder transforms the input waveform into a sequence of latent vectors that capture fine-grained acoustic details. These latent features are subsequently processed by the contextual module, which employs self-attention mechanisms to

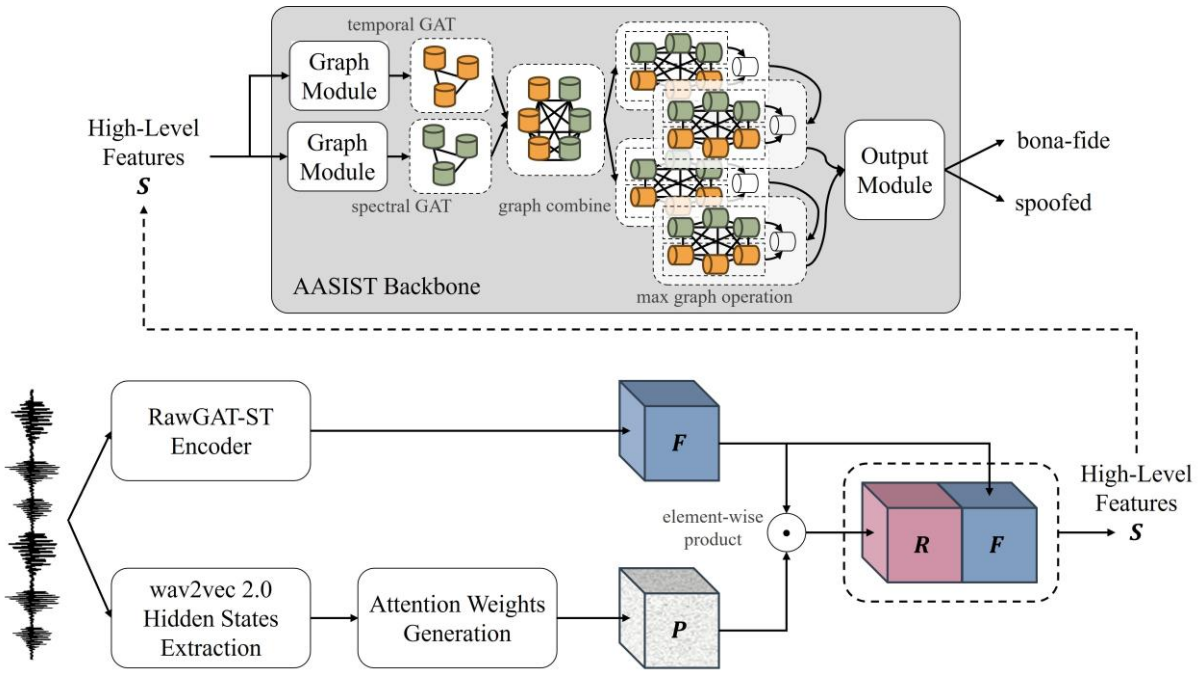


Figure 1: The proposed wav2vec 2.0-based attention network for high-level feature extraction in voice spoofing detection.

model temporal dependencies across the sequence, resulting in contextualized embeddings that reflect both short- and long-range speech characteristics. The model is pretrained using a contrastive objective, where segments of the latent sequence are masked and the network learns to distinguish the true representation from a set of distractors based on surrounding context. This training strategy enables wav2vec 2.0 to acquire phonetic and semantic knowledge from unlabeled speech data, making the learned representations broadly applicable to downstream tasks such as automatic speech recognition, speaker verification, and spoofing detection.

3 Speech Rule Generation via wav2vec 2.0-Based Attention

Unlike previous studies that directly utilize wav2vec 2.0 features as input to classification models, this work explores the use of wav2vec 2.0 representations to learn the underlying speech rules present in bona-fide speech. We hypothesize that spoofed speech introduces inconsistencies or deviations from these learned regularities. By identifying such rule violations, the proposed approach aims to enhance the accuracy of voice spoofing detection. The proposed wav2vec 2.0-based attention network for high-level feature extraction as depicted in Figure 1.

3.1 wav2vec 2.0-Based Attention

Initially, wav2vec 2.0 is used to extract hidden states s from the raw training audio, where $s \in \mathbb{R}^{T\times L}$ denotes a sequence of T time steps, each represented by an L-dimensional feature vector. These representations are then passed through N Residual Blocks for feature transformation. Assuming the $p_0 = s$, the standard transformation performed by each Residual Block is defined as follows:

$$p_{i+1} = \mathcal{F}(\mathcal{F}(p_i; \mathcal{K}_{i1}); \mathcal{K}_{i2}) + p_i \quad (1)$$

where $\mathcal{F}(\cdot)$ denotes a 2D convolutional layer parameterized by kernel $\mathcal{K}_{(\cdot)}$, and each input of $\mathcal{F}(\cdot)$ is nonlinearly transformed by a composite function consisting of batch normalization followed by the scaled exponential linear unit (SELU) activation.

After that, the feature size of p_N and the raw audio after encoder processing are different, we

Layer	Input shape	Output shape
Raw audio	-	(64600)
Wa2vec 2.0 hidden states	(64600)	(199, 768)
Expand dim	(199, 768)	(1, 199, 768)
ResBlock_A × 2	(1, 199, 768)	(C_1, 199, 768)
ResBlock_B × 4	(C_1, 199, 768)	(C_2, 199, 768)
Conv2D	(C_2, 199, 768) kernel: (7, 11) stride: (7, 11)	(C_2, 29, 69)
BN	-	-
MaxPool	(C_2, 29, 69) kernel: (1, 3)	(C_2, 29, 23)
Softmax	dim=1	(C_2, 29, 23)
Hybrid High-Level Features	(C_2, 29, 23) combine R, F	(C_2, 58, 23)

Table 1: The speech rule generation architecture for voice spoofing detection.

apply local convolution and max pooling to compress the size of p_{N} to match the encoder output size.

$$\mathcal{T} = \mathrm{MaxPool}(\mathrm{BN}(\mathcal{F}(p_{\mathrm{N}}))) \tag{2}$$

$$P_{c,t,f} = \frac{e^{\mathcal{T}_{c,t,f}}}{\sum_{\tau=1}^{T} e^{\mathcal{T}_{c,\tau,f}}} \tag{3}$$

where $\mathrm{MaxPool}(\cdot)$ is max pooling, $\mathrm{BN}(\cdot)$ is batch normalization, and $P \in \mathbb{R}^{C_2 \times T \times F}$ can be defined as the attention weights employed to regulate speech rules.

3.2 Hybrid High-Level Features

Since P is the attention weights derived from the hidden states of wav2vec 2.0, we further apply an element-wise product between P and the encoder output F to generate the corresponding speech rules.

$$R = P \odot F \tag{4}$$

Finally, the speech rule R is used as auxiliary features and concatenated with F to obtain the hybrid high-level features. The size for each layer is illustrated in Table 1.

4 Experimental Results

4.1 Data Preparation

In alignment with the data preparation methodology outlined in (J.-w. Jung et al., 2022), all experiments in this study are conducted using the LA partition of the ASVspoof 2019 dataset (M. Todisco et al., 2019). The dataset is divided into three distinct subsets: training, development, and evaluation. The training and development subsets include spoofed speech generated using six known attack algorithms (A01–A06), while the evaluation subset extends this with an additional set of seven attack methods (A07–A19). Furthermore, the ASVspoof 2021 (J. Yamagishi et al., 2021) evaluation set is used to evaluate the cross-corpus performance of the proposed voice spoofing detection method.

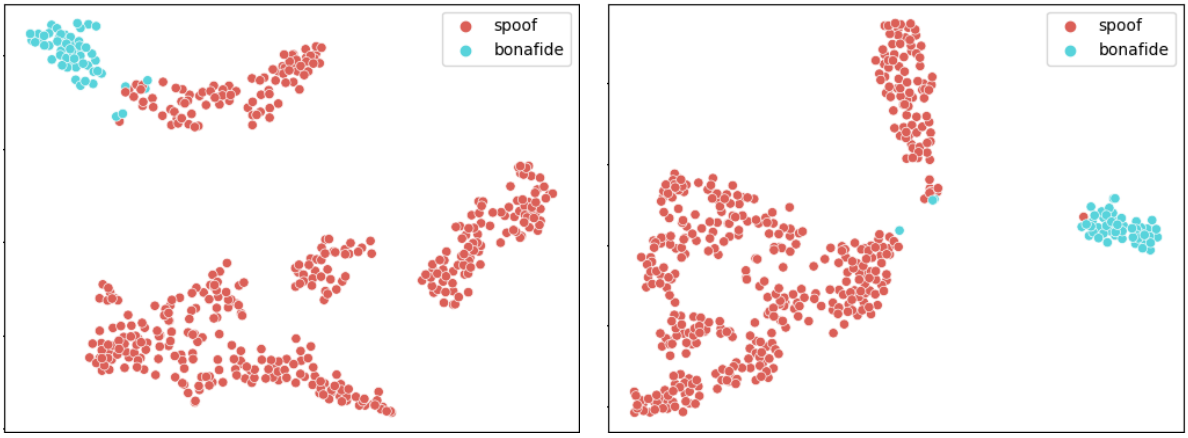
In this paper, we employ the "facebook/wav2vec2-base-960h" model, a Transformer-based architecture designed for speech representation learning. The model is pretrained in a self-supervised manner on 960 hours of unlabelled audio from the LibriSpeech corpus and later fine-tuned for automatic speech recognition tasks. Its structure consists of a convolutional feature extractor followed by twelve Transformer encoder layers, enabling the model to capture hierarchical representations of speech signals. We extract hidden states from both intermediate layers and the final layer. The intermediate layers are known to preserve more acoustic-level and phonetic information, making them well-suited for tasks that require detailed speech characteristics such as prosody, speaker traits, or subtle temporal variations. In contrast, the final layer tends to encode high-level semantic features aligned with the ASR objective, capturing more abstract linguistic content but potentially discarding lower-level acoustic cues.

4.2 Experimental Setup

In our experiments, we adopt lightweight variant AASIST-L as the backbone architecture, following the experimental setup outlined in (J.-w. Jung et al., 2022). The input to the model consists of raw audio waveforms with a fixed length of 64,600 samples, corresponding to approximately four seconds of speech. No data augmentation techniques are applied during training, ensuring that all models are trained on the original waveform data without synthetic variation. Model training is conducted

System	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	EER	min t-DCF
RawNet2 [13]	9.8	17.9	7.3	8.9	4.2	8.8	2.0	1.3	7.3	4.6	2.4	62.9	5.8	5.54	0.1547
RawGAT-ST [14]	1.19	0.33	0.03	1.54	0.41	1.54	0.14	0.14	1.03	0.67	1.44	3.22	**0.62**	1.19	0.0333
AASIST-L (reproduced)	**0.45**	0.34	0.02	**0.63**	0.34	**0.69**	0.19	0.23	0.53	**0.42**	1.96	2.97	0.88	1.14	0.0316
RULE-AASIST-L (proposed)	0.77	**0.16**	**0.02**	0.90	**0.16**	0.79	**0.12**	**0.10**	**0.42**	0.57	**1.18**	**2.34**	0.87	**0.86**	**0.0282**

Table 2: EER (%) and minimum t-DCF results for baseline and proposed model on the ASVspoof 2019 LA evaluation set.



(A) AASIST-L (B) RULE-AASIST-L

Figure 2: The distribution of output features from the last hidden layer of different models visualized using t-SNE, based on 240 randomly selected samples.

using the Adam optimizer with a batch size of 24 and a total of 100 training epochs. The objective function used is categorical cross-entropy loss.

As demonstrated in the study by (X. Wang and J. Yamagishi, 2021), the performance of spoofing detection systems can vary considerably depending on the choice of random seed due to the inherent stochasticity of the training process. To ensure a fair and robust evaluation, all experiments in this work are conducted using three different random seeds. In the experimental analysis, this paper reports the best result obtained from model training conducted with three different random seeds.

To evaluate system performance, we adopt two widely used metrics: the minimum tandem detection cost function (min t-DCF) and the equal error rate (EER).

4.3 Voice Spoofing Detection Results

The results are summarized in Table 2. Compared to the baseline systems, the proposed RULE-AASIST-L demonstrates significantly improved performance. Under the same backbone architecture and experimental setup, RULE-AASIST-L achieves a relative improvement of 24.6% in EER (i.e., 0.86% vs. 1.14%) and an 10.8% reduction in min t-DCF (i.e., 0.0282 vs. 0.0316), highlighting the effectiveness of the proposed method. On the other hand, Figure 2 illustrates the distribution of the output features from the last hidden layer of different models. It is evident that the proposed RULE-AASIST-L model yields more compact distributions for both spoof and bonafide classes compared to the AASIST-L baseline. This indicates that the RULE-AASIST-L model can more effectively distinguish between genuine and spoofed speech.

The results indicate that RULE-AASIST-L successfully leverages the attention generated during model training to define bona-fide speech rules. These learned rules help identify inconsistencies in spoofed speech, thereby

System	EER	min t-DCF
RULE-AASIST-L	**0.86**	**0.0282**
w/o ***F*** in the high-level features ***S***	1.54	0.0468
Use only ***P*** as the high-level features ***S***	2.80	0.0830
Replace wav2vec 2.0 hidden state extraction from layer 6 with layer 12	1.29	0.0371

Table 3: Results for ablation studies on AASIST-L backbone.

System	ASVspoof 2021 evaluation set	
	EER	min t-DCF
AASIST-L	13.65	0.4574
RULE-AASIST-L	**12.91**	**0.4347**

Table 4: EER (%) and minimum t-DCF results for baseline and proposed model on the ASVspoof 2021 LA evaluation set.

enhancing the system's voice spoofing detection capabilities.

Notably, the proposed approach does not directly use the wav2vec 2.0 features as input to the spoofing detection model. Instead, it employs these representations to construct speech rules, which in turn modulate the output of high-level features F. This indirect usage of wav2vec 2.0 features contributes to the strong performance gains observed. As a result, the method opens promising directions for future research on using self-supervised representations to guide rule-based structures in voice spoofing detection.

4.4 Ablation Study

Table 3 presents the results of ablation experiments, in which individual components of the AASIST model are either removed or replaced. The results show a clear drop in performance when only the speech rule R is used as the high-level representation S. This performance degradation is attributed to the fact that R, while effective in modeling sequential consistency, lacks the rich acoustic information contained in the original high-level features F, making it insufficient on its own for effective spoofing detection. Similarly, replacing S directly with attention weights P results in an even more significant decline in performance. This suggests that attention weights alone, without the support of learned feature representations, are inadequate as standalone features.

Finally, we examine the effect of changing the source layer for feature extraction within the wav2vec 2.0 encoder. When the hidden states are extracted from layer 12 (the final layer) instead of layer 6 (an intermediate layer), a noticeable performance drop is observed. This can be explained by the representational nature of the final layer, which is optimized for ASR and tends to encode more abstract semantic features. While such features are useful for linguistic understanding, they often lack the lower-level acoustic cues that are critical for spoofing detection, thereby reducing detection effectiveness.

4.5 Cross-Corpus Evaluation

In this experiment, the ASVspoof 2021 LA evaluation set was further used to evaluate the cross-corpus performance of voice spoofing detection as shown in Table 4. It is evident that training solely on the ASVspoof 2019 training set and evaluating on the ASVspoof 2021 evaluation set leads to an increase in EER due to data mismatch. Nevertheless, the proposed RULE-AASIST-L model consistently outperforms the baseline AASIST-L, demonstrating that the wav2vec 2.0-based attention mechanism remains effective in improving the performance of voice spoofing detection in cross-corpus evaluations.

5 Conclusions

This work introduces RULE-AASIST-L, a rule-aware voice spoofing detection framework that utilizes attention-derived speech rules based on wav2vec 2.0 representations. By modeling the correlation between acoustic features and attention weights, the proposed method captures rule-based inconsistencies introduced by synthetic speech. Unlike previous methods that treat wav2vec 2.0 features as direct inputs, our approach exploits these representations to guide the learning of bona-fide speech patterns, thereby improving detection robustness. Experimental results on the ASVspoof 2019 LA dataset confirm the effectiveness of our method, with substantial performance gains over

baseline systems. Ablation experiments further underscore the importance of rule modeling and the choice of representation layer, showing that intermediate-layer features (e.g., layer 6) retain richer acoustic cues than final-layer representations. In the future, this study opens new directions for integrating self-supervised learning and rule-based reasoning in the field of voice spoofing detection, and we plan to further investigate the possibility of utilizing the constructed speech rules during the inference stage without relying on wav2vec 2.0 features. One potential direction involves integrating alignment search and a flow-based module to generate approximated wav2vec 2.0 representations during inference, thereby eliminating the need for direct feature extraction from the original model.

Acknowledgments

This work was supported in part by the National Science and Technology Council (NSTC), Taiwan, under Grant No. NSTC 113-2222-E-218 -003 -MY2.

References

Y. Zhang, F. Jiang, and Z. Duan. 2021. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters, Volume 28*, pages 937–941.

J. Zhou, T. Hai, D. N. A. Jawawi, D. Wang, E. Ibeke, and C. Biamba. 2022. Voice spoofing countermeasure for voice replay attacks using deep learning. *Journal of Cloud Computing*, 11:51.

A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez. 2019. A light convolutional GRU-rnn deep feature extractor for ASV spoofing detection. In *Proceedings of INTERSPEECH*, pages 1068–1072.

J. Boyd, M. Fahim, and O. Olukoya. 2023. Voice spoofing detection for multiclass attack classification using deep learning. *Machine Learning With Applications*, 14:100503.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations," In *Proceedings of Neural Information Processing Systems (NeurIPS)*, pages 12449–12460.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing, Volume 29*, pages 3451–3460.

S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing, Volume 16, Number 6*, pages 1505–1518.

A. Bawitlung, S. K. Dash, and R. M. Pattanayak. 2025. Mizo Automatic Speech Recognition: Leveraging Wav2vec 2.0 and XLS-R for Enhanced Accuracy in Low-Resource Language Processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Z. Fan, M. Li, S. Zhou, and B. Xu. 2021. Exploring wav2vec 2.0 on speaker verification and language identification. In *Proceedings of INTERSPEECH*, pages 1509–1513.

B. Nasersharif and M. Namvarpour. 2024. Exploring the potential of Wav2vec 2.0 for speech emotion recognition using classifier combination and attention-based feature fusion. *The Journal of Supercomputing, Volume 80, Number 16*, pages 23667–23688.

H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *The Speaker and Language Recognition Workshop (Odyssey)*, pages 112–119.

H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans. 2021. End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*.

J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans. 2022. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 6367–6371.

M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee. 2019. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In *Proceedings of INTERSPEECH*, pages 1008–1012.

J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 47–54.

X. Wang and J. Yamagishi. 2021. A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection. In *Proceedings of INTERSPEECH*, pages 4259–4263.