Fusion of Modulation Spectrogram and SSL with Multi-head Attention for Fake Speech Detection

Rishith Sadashiv T N[1] Abhishek Bedge[1] Saisha Suresh Bore[2] Jagabandhu Mishra[3]
Mrinmoy Bhattacharjee[4] S R Mahadeva Prasanna[1,2]
[1]IIT Dharwad [2]IIIT Dharwad [3]University of Eastern Finland [4]IIT Jammu
Email: ee24dpp10@iitdh.ac.in

***Abstract*—Fake speech detection systems have become a necessity to combat against speech deepfakes. Current systems exhibit poor generalizability on out-of-domain speech samples due to lack to diverse training data. In this paper, we attempt to address domain generalization issue by proposing a novel speech representation using self-supervised (SSL) speech embeddings and the Modulation Spectrogram (MS) feature. A fusion strategy is used to combine both speech representations to introduce a new front-end for the classification task. The proposed SSL+MS fusion representation is passed to the AASIST back-end network. Experiments are conducted on monolingual and multilingual fake speech datasets to evaluate the efficacy of the proposed model architecture in cross-dataset and multilingual cases. The proposed model achieves a relative performance improvement of 37% and 20% on the ASVspoof 2019 and MLAAD datasets, respectively, in in-domain settings compared to the baseline. In the out-of-domain scenario, the model trained on ASVspoof 2019 shows a 36% relative improvement when evaluated on the MLAAD dataset. Across all evaluated languages, the proposed model consistently outperforms the baseline, indicating enhanced domain generalization.**

I. Introduction

In recent years, the sophistication of machine-generated speech has increased significantly, enabling both beneficial and malicious applications. While generative speech technology supports valuable use cases such as assistive tools and accessibility, it also poses serious threats when misused—for instance, in spreading manipulated war narratives or deceiving speaker verification (SV) systems. The rapid progress in this field introduces ongoing challenges for designing effective countermeasure systems, particularly for *Fake Speech Detection (FSD)*. FSD has been extensively studied, evolving from the use of hand-crafted features and simple classifiers to end-to-end deep neural networks like RawNet2 [1] and AASIST [2]. More recently, SSL models and state-space architectures like Mamba have shown promise [3], [4]. For real-time deployment, systems trained in one domain must generalize well to others. However, domain generalizability remains a persistent challenge, as FSD models often struggle to maintain performance across datasets due to variations in recording conditions and dataset-specific characteristics.

Many works have established the performance degradation of FSD systems in out-of-domain scenarios [5], [6]. To resolve this, the attempts are broadly in two directions to improve generalization ability of FSD system, (1) use of specialized model training strategies, and (2) signal processing and data-driven approaches. Various training strategies have been explored to improve generalization, including multi-task meta-learning [7], continual learning [8], one-class learning [9], and optimal transport-based domain adaptation [10]. Many existing approaches focus on signal processing techniques to extract novel features for FSD. For instance, the study in [11] proposed the application of 2D Discrete Cosine Transform (2D-DCT) on log-Mel spectrograms. Pronunciation and prosodic features have also been explored in [12] to enhance generalization. Furthermore, a combination of modulation spectrogram and residual modulation spectrogram features has been investigated in [13]. SSL front-ends have gained popularity in recent years. The study in [14] demonstrated that fine-tuning the wav2vec 2.0 XLS-R model on an FSD dataset leads to improved domain generalization, even when paired with a simple fully connected (FC) back-end. Furthermore, the results indicate that the wav2vec 2.0 model provides better FSD generalization compared to other SSL models like HuBERT. Similarly, another work [15] investigates the use of a variational information bottleneck module along with a wav2vec-based front-end and an FC back-end. However, we hypothesize that a representation derived by combining signal processing-based features with data-driven SSL embeddings could potentially be a promising approach for the FSD task.

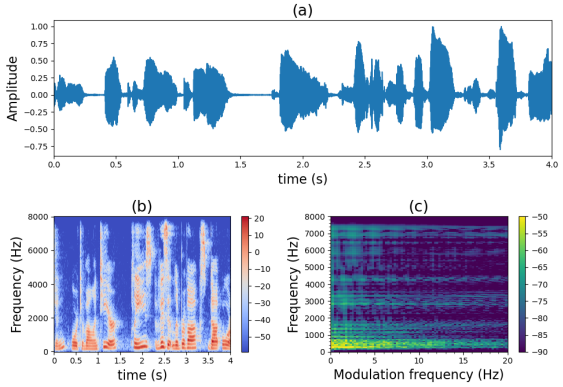

Fig. 1. (a) Speech signal, (b) Spectrogram, and (c) Modulation spectrogram

In this study, we propose a *novel front-end representation for improved domain generalization* in the FSD task. We achieve this by *combining* wav2vec 2.0 cross-lingual self-supervised speech representations (*XLS-R*), which is hereafter

979-8-3315-7206-8/25/$31.00

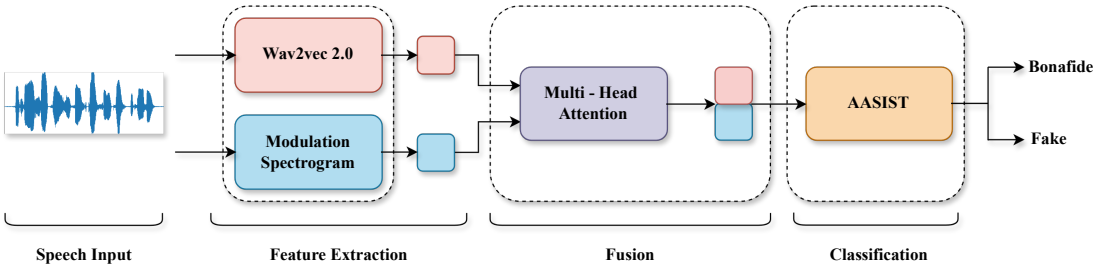

Fig. 2. Proposed methodology: In the fusion block, wav2vec 2.0 embeddings form the key and value, with the modulation spectrogram as the query.

referred to as the SSL model, with the *modulation spectrogram* feature. While the modulation spectrogram has been previously introduced for FSD in [13], and SSL model embeddings have been combined with other speech features to enhance generalizability in [12], to the best of our knowledge, the joint use of modulation spectrogram and SSL embeddings has not yet been explored. To address this gap, we employ a multi-head attention mechanism as the fusion strategy. Since SSL models are primarily trained to capture speech characteristics at the word or syllable level, they may not effectively represent frame-level artifacts. In contrast, modulation spectrogram provides variation in speech dynamics from frame-level to prosodic level. Fig. 1 illustrates the modulation spectrogram feature alongside the corresponding speech waveform and spectrogram. We hypothesize that the fusion of SSL embeddings with modulation spectrogram features can yield a more generalizable representation for FSD. The AASIST network is employed as the back-end architecture, as described in [3]. The effectiveness of the proposed system is evaluated on the monolingual ASVspoof 2019 Logical Access (LA) dataset, followed by domain generalization experiments using the recent multilingual MLAAD dataset [16]. Additionally, the impact of language variation is analyzed using the MLAAD dataset. Experimental results demonstrate that the proposed fusion-based front-end significantly enhances domain generalization compared to the baseline. The main contributions of this paper are summarized as follows:

- We propose the fusion of modulation spectrogram feature with SSL model embeddings for the FSD task.
- A novel architecture is introduced that employs the fused feature representation as the front-end and the AASIST network as the back-end.
- Validation of the proposed framework on cross-domain and multi-lingual setup.

The remainder of the paper is organized as follows: Section II illustrates the proposed methodology. The experimental setup is described in Section III. The results are reported in Section IV along with discussions. The paper is concluded in Section V.

II. Methodology

In this section, we describe our proposed approach for the FSD task, which fuses SSL features with modulation spectrogram using multi-head attention. Fig. 2 illustrates the overall architecture. Building on the widespread use of Audio Anti-Spoofing using Intgrated Spectro-Temporal graph attention networks (AASIST) with SSL features in prior work [2], [3], we combine the fused representation with AASIST to perform spoofing detection. In the following subsections, we briefly explain the modulation spectrogram, SSL features, the fusion process using multi-head attention, and the AASIST model.

A. Modulation Spectrogram.

The modulation spectrogram provides a two-dimensional representation of a speech signal. To compute it, we follow a two-step process. First, we apply a Short-Term Fourier Transform (STFT) to the speech signal $x(t)$ to obtain the spectrogram $X(t, f)$, which serves as a time-frequency representation. The frequency $f \in [0, f_N]$ and time $t \in [0, T]$, where N and T denote the number of FFT points and the total number of time samples, respectively. Next, we compute the modulation spectrogram $Y(f_{mod}, f)$ by applying a Fourier transform over time to the magnitude of each frequency component of $X(t, f)$. This transformation yields:

$$Y(f_{mod}, f_i) = \mathcal{F}\{|X(t, f_i)|\}, i = 0 \dots f_N \quad (1)$$

where $\mathcal{F}$ denotes the Fourier transform. We use $\hat{N}$ FFT points for modulation spectrogram computation, which is equal to the number of frames in the STFT. The resulting modulation spectrogram captures the conventional frequency f along one axis and the modulation frequency f_{mod} along the other [17]. The modulation frequency f_{mod} captures how the temporal dynamics of the speech signal vary, from rapid changes at the frame level to slower trends at the prosodic level.

B. SSL Embeddings

We use the XLS-R variant of the wav2vec 2.0 model [18], [19] to extract feature representations from speech signals. The model employs a multi-layer convolutional neural network as a feature encoder $f : \mathcal{X} \rightarrow \mathcal{Z}$, which transforms raw waveforms $x_{1:T}$ into latent speech representations $z_{1:\hat{T}}$ where T denotes the number of samples and $\hat{T}$ represents the number of time steps. The stride of the feature encoder determines the value of $\hat{T}$. The model then feeds the speech representations into transformer network $g : \mathcal{Z} \rightarrow \mathcal{C}$, which produces context representations $c_{1:\hat{T}}$ that capture information from the entire latent representation sequence in an end-to-end manner.

During self-supervised training, the latent speech representations $z_{1:\hat{T}}$ are quantized to a finite set of speech representations $q_{1:\hat{T}}$ using a quantization module $\mathcal{Z} \rightarrow \mathcal{Q}$. It involves

quantized representations from multiple codebooks. The latent representations are masked at random starting points before being fed to the transformer network. The model training involves solving a contrastive task, which requires identifying the true quantized latent representation q_t for a masked time step within a set of distractors. The contrastive loss is augmented with a codebook diversity loss to encourage the model to use all codebook entries. The XLS-R model with 0.3 billion parameters available at Fairseq toolkit [20] is used in our study.

C. Fusion Strategy using Multi-Head Attention

We perform the fusion of XLS-R embeddings and modulation spectrogram using a multi-head attention network [12]. The multi-head attention mechanism conducts multiple scaled dot-product attention operations, as defined in (2):

$$Attention(Q, K, V) = Softmax\left(\frac{KQ^T}{\sqrt{d_k}}\right)V, \quad (2)$$

where we use the XLS-R embeddings as the key (K) and value (V), and the modulation spectrogram feature as the query (Q). The key and query both have dimensionality d_k, and the value has dimensionality d_v. We perform projection operations using multiple FC layers to generate h sets of (Q, K, V) representations. We apply the attention operation to each set in parallel. Then, we concatenate the resulting outputs from all attention heads and project them through an FC layer [21], as shown in,

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, \cdots, head_h)W^O \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \quad (3)$$

where W_i^Q, W_i^K, W_i^V, and W_i^O denote the parameter matrices of the FC layers for the i^{th} head, and $i \in [1, h]$.

D. AASIST Spoofing Detection

AASIST is a widely used graph neural network framework for FSD. It uses a sinc-convolution layer to extract front-end features from raw audio, which are then encoded by a RawNet2 variant [1]. The model reshapes the output into a 2D representation and passes it through six residual blocks to extract high-level features. Two parallel graph modules, each with graph attention and pooling layers, model spectral and temporal artifacts. A max graph operation combines their outputs using two heterogeneous graph branches, followed by an element-wise maximum. Each branch uses two HS-GAL layers and pooling, with a stack node aggregating information. Final output uses max and average pooling, followed by a two-node output layer [2]. In [3], wav2vec 2.0 replaces the sinc-convolution layer, and its embeddings are input to RawNet2. In our method, we replace wav2vec 2.0 with fused features, which are passed to the AASIST back-end.

III. Experimental Setup

A. Datasets

We use three datasets in this work: ASVspoof 2019, ASVspoof 2021, and MLAAD. We choose ASVspoof 2019 due to its widespread use in the literature, ASVspoof 2021 to assess generalizability under channel and noise variations, and MLAAD to evaluate generalization across different languages. Table I summarizes the key characteristics of these datasets, and the following subsections provide a brief description of each.

TABLE I
Composition of the datasets

Dataset		ASVspoof 2019 LA	ASVspoof 2021 LA	MLAAD
Train	Bonafide	2580	-	28345
	Fake	22800	-	36566
Development	Bonafide	2548	-	6584
	Fake	22296	-	9765
Evaluation	Bonafide	7355	14816	6390
	Fake	63882	133360	19675

1) ASVspoof 2019: We use the LA partition of the ASVspoof 2019 dataset, which includes fake speech samples generated using 17 different neural acoustic and waveform-based TTS and VC spoofing techniques. The dataset is divided into three subsets: train, development, and evaluation, each containing non-overlapping speakers. The train and development sets include fake speech samples from 6 known spoofing techniques, while the evaluation set contains samples from 13 spoofing techniques, 2 known and 11 unknown. The bonafide (genuine) speech samples come from the VCTK corpus [22]. This dataset is monolingual and includes only English-language speech.

2) ASVspoof 2021: The ASVspoof 2021 dataset provides an updated evaluation set with an increased number of bonafide and fake speech samples. Unlike ASVspoof 2019, which contains studio-quality recordings, the 2021 evaluation samples are passed through telephony systems (VoIP and PSTN) to simulate real-world, in-the-wild conditions [23]. This dataset is also monolingual and contains only English-language speech.

3) MLAAD: The MLAAD dataset contains fake speech samples generated in 23 different languages using 52 state-of-the-art models across 22 architectures [16]. It builds on the M-AILABS speech dataset [24], which provides bonafide speech in 8 European languages. For languages not covered by M-AILABS, English text is translated into the target languages and then synthesized into fake speech using various TTS models sourced from *Coqui.ai*[1] and *Hugging Face*. Following the protocols from [25], we split the dataset into training, development, and evaluation subsets with no speaker overlap. The bonafide samples come from 5, 4, and 4 languages in the train, development, and evaluation sets respectively, while the fake samples include all 23 languages across each subset.

B. Evaluation Metric

We use Equal Error Rate (EER) as the metric throughout this work. It represents the threshold at which the false alarm rate and miss rate are approximately equal, as shown in (4).

$$EER = P_{fa}^{cm}(\tau_{EER}) \approx P_{miss}^{cm}(\tau_{EER}) \quad (4)$$

[1]https://github.com/coqui-ai/TTS

C. Modulation Spectrogram Extraction

We restrict all audio samples in this work to approximately 4 seconds with a sampling rate of 16 kHz ($64,600$ samples). We zero-pad shorter audios to match this length. Then, we extract the modulation spectrogram feature from the speech signal using a frame length of 25 ms and a frame shift of 10 ms. For STFT computation, the number of FFT points is set equal to the window length, i.e., $0.025 \times 16000 = 400$. For modulation spectrogram computation, the number of FFT points ($\hat{N}$) is set to the number of STFT frames, i.e., 402. This results in a modulation spectrogram feature dimension of $(\frac{N}{2}+1) \times (\frac{\hat{N}}{2}+1) = 201 \times 202$.

D. Fusion using multi-head attention

The fusion operation follows a similar approach to that described in [12]. The self-supervised XLS-R (0.3B) model[2] is used as SSL model. Raw audio segments of approximately 4 seconds are input to the SSL model, producing embeddings of size 201×1024. These embeddings are subsequently projected to a lower dimension of 201×128 via an FC layer. For the fusion strategy, the key and value representations are obtained by passing the SSL embeddings through two separate FC layers, while the query is derived from the modulation spectrogram feature using another FC layer. These components are then processed by a multi-head attention block with $h = 4$ heads. The output of the attention block is further projected through a final FC layer. All FC layers used in the fusion module output a fixed dimension of $P = 256$, resulting in a final fused representation of size $201 \times P$. This fusion representation is then fed into the AASIST back-end network. The entire model, including the SSL front-end, fusion module, and back-end, is jointly optimized during training.

E. Training Details

We apply RawBoost data augmentation on the fly to the existing training data using the same parameters and configuration as the baseline work [3]. We use a batch size of 14 and a fixed learning rate of 10^{-6}. We optimize the model with the standard Adam optimizer and use a weighted cross-entropy loss. All models are trained for 100 epochs on an A100 GPU and we choose the model with the best development loss for testing. The implementation is available in the Github repo[3].

IV. Results and Discussions

This section reports the results of the baseline and proposed fusion models. We use SSL-AASIST [3] as the baseline and denote the proposed fusion model of SSL and modulation spectrogram as (SSL+MS)-AASIST. We train models on ASVspoof 2019, MLAAD, and their combination, and evaluate them on the ASVspoof 2019, ASVspoof 2021, and MLAAD evaluation sets. Table II presents the performance across all evaluation scenarios.

[2]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec
[3]https://github.com/rishithSadashiv/ssl-ms-fsd

TABLE II
Cross-dataset EER (%) comparison between the baseline (SSL) and the proposed fusion (SSL+MS) architectures. The LA part of the ASVspoof datasets has been used. Combined denotes ASVspoof 2019 + MLAAD datasets.

Model	Train Set ↓	Test Set → ASVspoof 2019	ASVspoof 2021	MLAAD
SSL	ASVspoof 2019	0.27	1.02	27.97
	MLAAD	38.49	37.85	8.24
	Combined	1.33	15.09	9.72
SSL+MS	ASVspoof 2019	0.17	1.15	17.89
	MLAAD	40.89	48.45	6.52
	Combined	0.34	3.04	5.79

A. Baseline: SSL with AASIST

The baseline model trained on the ASVspoof 2019 dataset achieves strong in-domain performance with an EER of 0.27% and generalizes reasonably well to the ASVspoof 2021 dataset, where it reaches an EER of 1.02%. However, it performs poorly on the out-of-domain MLAAD dataset, yielding a high EER of 27.97%. When trained on the MLAAD dataset, the model records an in-domain EER of 8.24% but fails to generalize, with EERs of 38.49% on ASVspoof 2019 and 37.85% on ASVspoof 2021. In comparision to the in-domain performance, training the model on the combined ASVspoof 2019 and MLAAD datasets leads to a slight degradation in ASVspoof 2019 performance (EER of 1.33%), a substantial drop in ASVspoof 2021 performance (EER of 15.09%), and a moderate decrease on MLAAD (EER of 9.72%). These results show that although the baseline model performs well in in-domain settings, it struggles to generalize across domains, highlighting the impact of dataset-specific characteristics on model performance.

B. Proposed: Fusion of Modulation spectrogram and SSL embeddings with AASIST

We conducted the same set of cross-domain experiments using the proposed fusion model. Notably, the fusion model achieves improved in-domain results, with an EER of 0.17% on ASVspoof 2019 and 6.52% on MLAAD, outperforming the corresponding baseline models. In out-of-domain evaluations, the ASVspoof 2019-trained fusion model performs comparably to its baseline counterpart on the ASVspoof 2021 dataset and shows enhanced performance on MLAAD. In contrast, the MLAAD-trained fusion model continues to perform poorly on both ASVspoof datasets, mirroring the trend observed in the baseline. This may be attributed to insufficient number of English language speech samples in MLAAD dataset.

Interestingly, the fusion model trained on the combined ASVspoof 2019 and MLAAD datasets demonstrates significant improvements. While its performance on ASVspoof 2019 and ASVspoof 2021 slightly lags behind the ASVspoof 2019-only trained fusion model, it clearly outperforms the baseline across all evaluation sets. On the MLAAD dataset, it even surpasses the in-domain performance of the MLAAD-trained fusion model. These results suggest that incorporating diverse data during training enables the fusion model to learn broader

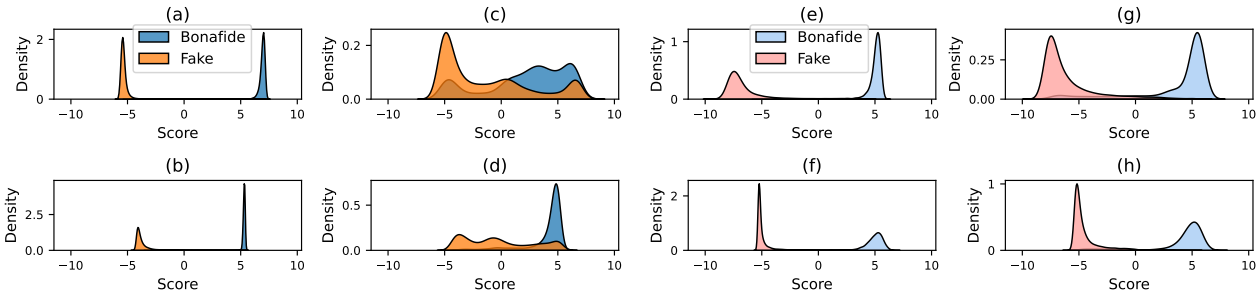

Fig. 3. Score density plots for bonafide and fake speech across different datasets. Plots (a)–(d) show models trained on the ASVspoof 2019 dataset, while plots (e)–(h) show models trained on the combined ASVspoof 2019 and MLAAD datasets. (a) Baseline model on ASVspoof 2019 eval set (0.27% EER), (b) Fusion model on ASVspoof 2019 eval set (0.17% EER), (c) Baseline model on MLAAD eval set (27.97% EER), (d) Fusion model on MLAAD eval set (17.89% EER), (e) Baseline model on ASVspoof 2019 eval set (1.33% EER), (f) Fusion model on ASVspoof 2019 eval set (0.34% EER), (g) Baseline model on MLAAD eval set (9.72% EER), (h) Fusion model on MLAAD eval set (5.79% EER).

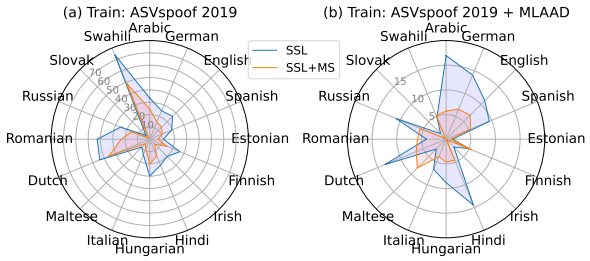

Fig. 4. Radar plots of EER (%) for baseline (SSL) and proposed (SSL+MS) models across languages in the MLAAD evaluation set. (a) Out-of-domain case: models are trained using monolingual ASVspoof 2019 LA dataset (Overall performance: SSL: 27.97% EER and SSL+MS: 17.89% EER). (b) Models are trained on combination of ASVspoof 2019 LA + MLAAD dataset (Overall performance: SSL: 9.72% EER and 5.79% EER).

feature representations, which improves its generalizability and robustness to domain shifts.

Fig. 3 presents density plots of classification scores from different models across various evaluation sets. The top row corresponds to the baseline SSL model, and the bottom row represents the proposed fusion (SSL+MS) model. Plots (a) and (b), which show in-domain results on ASVspoof 2019, reveal clean score separation for both models. However, in the out-of-domain case (plots c and d), where models are trained on ASVspoof 2019 and evaluated on MLAAD, the fusion model shows narrower bonafide score distribution, indicating improved domain generalization despite high EER values (27.97% for baseline vs. 17.89% for fusion).

In the last four plots (e–h), we repeat the experiment using the combined training set. Comparing these plots clearly shows that the fusion model consistently achieves better separation between bonafide and fake scores than the baseline, reflecting the EER trends reported in Table II. In summary, the proposed fusion architecture not only enhances in-domain performance but also significantly improves generalization across domains. By leveraging the additive information from modulation spectrograms and SSL embeddings, the fusion model demonstrates robustness to dataset variations and offers a promising direction toward generalizable fake speech detection.

C. Generalization across language

We analyze the behavior of both the baseline and proposed fusion models across individual languages in the MLAAD evaluation set, under two training conditions: using only the ASVspoof 2019 dataset (monolingual English) and using the combined ASVspoof 2019 and MLAAD datasets (multilingual). The evaluation protocol includes bonafide speech from four languages—German, Spanish, Russian, and Ukrainian—and fake speech across 23 languages. For each fake language, we compute the EER using its scores as false and all bonafide scores as true. Fig. 4 presents these language-wise EERs in a radar chart, excluding seven languages with fewer than 100 fake samples.

Plot (a) shows the results for models trained on ASVspoof 2019, where the proposed fusion model consistently outperforms the baseline across all evaluated languages, suggesting better generalization in out-of-domain scenarios. Plot (b) presents the performance when models are trained on the combined ASVspoof 2019 and MLAAD datasets. The overall EERs are lower than in plot (a), indicating the benefits of multilingual training. The fusion model continues to show improved results for most languages, with performance comparable to the baseline in Maltese, Finnish, and Romanian. These findings suggest that the proposed fusion model provides improved language robustness for fake speech detection, showing better generalization in both cross-lingual and multilingual training scenarios compared to the SSL-AASIST baseline.

V. Conclusion

This paper presents a novel approach for improving domain generalization in FSD by fusing modulation spectrogram feature with SSL embeddings. The proposed fusion leverages additive information providing a more generalizable representation. Integrated with the AASIST back-end, the (SSL+MS)-AASIST model outperforms the SSL-AASIST baseline in both in-domain and most out-of-domain evaluations. Additionally, the model demonstrates enhanced language robustness in multilingual scenarios. Future work will focus on exploring the integration of additional features and advanced training strategies for further performance improvement.

Acknowledgment

This work was supported by MeitY, RCI Hyderabad, and SERB, India through various projects. Jagabandhu Mishra was supported by Academy of Finland ("SPEECHFAKES").

REFERENCES

[1] 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.

[2] J.-w. Jung, H.-S. Heo, H. Tak, *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2022, pp. 6367–6371.

[3] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119. DOI: 10.21437/Odyssey.2022-16.

[4] Y. Xiao and R. K. Das, "XLSR-Mamba: A dual-column bidirectional state space model for spoofing attack detection," *IEEE Signal Processing Letters*, 2025.

[5] Y. Zhang, G. Zhu, F. Jiang, and Z. Duan, "An Empirical Study on Channel Effects for Synthetic Voice Spoofing Countermeasure Systems," in *Proc. Interspeech 2021*, 2021, pp. 4309–4313. DOI: 10.21437/Interspeech.2021-1820.

[6] 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.

[7] L. Wang, L. Yu, Y. Zhang, and H. Xie, "Generalizable speech spoofing detection against silence trimming with data augmentation and multi-task meta-learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[8] H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang, "Continual learning for fake audio detection," in *Interspeech 2021*, 2021, pp. 886–890. DOI: 10.21437/Interspeech.2021-794.

[9] G. Lin, W. Luo, D. Luo, and J. Huang, "One-class neural network with directed statistics pooling for spoofing speech detection," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2581–2593, 2024.

[10] R. Zhang, J. Wei, X. Lu, *et al.*, "SHDA: Sinkhorn domain attention for cross-domain audio anti-spoofing," *IEEE Transactions on Information Forensics and Security*, 2025.

[11] Y. Gao, T. Vuong, M. Elyasi, G. Bharaj, and R. Singh, "Generalized Spoofing Detection Inspired from Audio Generation Artifacts," in *Proc. Interspeech 2021*, 2021, pp. 4184–4188. DOI: 10.21437/Interspeech.2021-1705.

[12] C. Wang, J. Yi, J. Tao, C. Y. Zhang, S. Zhang, and X. Chen, "Detection of cross-dataset fake audio based on prosodic and pronunciation features," in *Interspeech 2023*, 2023, pp. 3844–3848. DOI: 10.21437/Interspeech.2023-1254.

[13] R. Sadashiv TN, D. Kumar, A. Agarwal, M. Tzudir, J. Mishra, and S. M. Prasanna, "Source and system-based modulation approach for fake speech detection," in *International Conference on Speech and Computer*, Springer, 2023, pp. 142–155.

[14] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 100–106. DOI: 10.21437/Odyssey.2022-14.

[15] Y. Eom, Y. Lee, J. S. Um, and H. R. Kim, "Anti-Spoofing Using Transfer Learning with Variational Information Bottleneck," in *Proc. Interspeech 2022*, 2022, pp. 3568–3572. DOI: 10.21437/Interspeech.2022-10200.

[16] N. M. Müller, P. Kawa, W. H. Choong, *et al.*, "MLAAD: The multi-language audio anti-spoofing dataset," in *2024 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2024, pp. 1–7.

[17] R. Cassani, I. Albuquerque, J. Monteiro, and T. H. Falk, "AMA: An open-source amplitude modulation analysis toolkit for signal processing applications," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE, 2019, pp. 1–4.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[19] A. Babu, C. Wang, A. Tjandra, *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[20] M. Ott, S. Edunov, A. Baevski, *et al.*, "fairseq: A fast, extensible toolkit for sequence modeling," *arXiv preprint arXiv:1904.01038*, 2019.

[21] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101 114, 2020.

[23] X. Liu, X. Wang, M. Sahidullah, *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[24] imdatceleste, "The M-AILABS speech dataset," *Github*, Jun. 17, 2025. [Online]. Available: https://github.com/imdatceleste/m-ailabs-dataset.

[25] N. Klein, T. Chen, H. Tak, R. Casal, and E. Khoury, "Source tracing of audio deepfake systems," in *Interspeech 2024*, 2024, pp. 1100–1104. DOI: 10.21437/Interspeech.2024-1283.