MSS: A Multilingual Spoofed Speech dataset with code-switching for anti-spoofing measures

*Muhammad Hamza, †Hafsa Ilyas, *Junaid Mir, †Ali Javed, ‡Muhammad Haroon Yousaf, §Ahmed Zoha
*Department of *Electrical Engineering, †Software Engineering and ‡Computer Engineering,*
University of Engineering and Technology Taxila, Pakistan
§*James Watt School of Engineering, University of Glasgow, United Kingdom*
*Emails: {*junaid.mir@, †ali.javed@, ‡haroon.yousaf@}uettaxila.edu.pk, §ahmed.zoha@glasgow.ac.uk*

***Abstract*—A significant proportion of the world's population speaks Urdu and Hindi, with many individuals being bilingual in both English and these languages. Still, no multilingual spoofing dataset exists to capture the conversational style of bilingual speakers who frequently code-switch while communicating. This paper presents a multilingual spoofed speech (MSS) dataset comprising 472,486 utterances from 154 speakers. We specifically considered bona fide utterances from Urdu and Hindi speakers, where language alternation occurs within a single audio. Spoofed samples are generated using voice conversion techniques to preserve the speaking accents and conversation styles of bilingual individuals. Further, we propose and evaluate an anti-spoofing framework called WavSpeech-AASIST, which incorporates self-supervised models (wav2vec and UniSpeech) into the AASIST network. Our comparative analysis underscores the significance of the MSS dataset and demonstrates the effectiveness of WavSpeech-AASIST for audio spoofing detection.**

***Index Terms*—Audio deepfakes, code switching, multilingual spoofing dataset, SSL, voice spoofing detection**

I. Introduction

Advancements in generative artificial intelligence (AI) algorithms have significantly evolved audio deepfake generation techniques, enabling them to synthesize a person's voice accurately by mimicking speech patterns, pitch, and timbre. Two main categories of speech synthesis (SS) techniques exist: text-to-speech (TTS) and voice conversion (VC). TTS algorithms primarily use vocoders to generate synthetic speech that adheres to the linguistic rules of the given text. In contrast, VC methods replicate a target speaker's voice while maintaining the content of the original speaker's speech [1]. With recent developments in SS techniques [2], voice spoofing technology has emerged as a considerable threat primarily due to its potential for misuse, such as spreading misinformation, which raises concerns about the integrity of multimedia content. Moreover, voice-enabled devices and automatic speaker verification (ASV) systems are at risk due to the generation of highly realistic spoofed voices.

The rise of audio spoofing threats has led to the introduction of various datasets aimed at developing spoofing countermeasures. One of the significant contributions to the collection of spoofed voice datasets for the English language is the ASVspoof challenge series, which biannually releases spoofed datasets in English to enhance the ASV systems. The ASVspoof-2015 [12] was the first edition of the ASV challenge, and the latest version is the ASVspoof-5 dataset [11], featuring spoofed samples generated using TTS, VC, and adversarial attacks. Additionally, several other datasets (as mentioned in Table I) have been proposed for languages other than English. However, these datasets do not include language alternation, and a speaker uses only a single selected language throughout his speech without code-switching. Moreover, most of the existing audio spoofing datasets are monolingual. These observations highlight a noticeable research gap in multilingual and code-switched spoofing datasets, underscoring the need for development in audio spoofing detection.

Hindi and Urdu, the lingua Francas of India and Pakistan, respectively, are the two major South Asian languages, with over half a billion speakers. Hindi is the third most widely spoken language in the world as of 2024. The English accent and conversation style among native Urdu and Hindi speakers vary across the South Asia region and are considerably different from those of native English speakers. It is common for Urdu and Hindi speakers to switch between Urdu and English or between Hindi and English during conversation. However, no spoofing dataset accurately represents the speaking style and conversation dynamics of Urdu and Hindi speakers, focusing primarily on Urdu and Hindi alongside the English language. Consequently, the performance of spoofing detection systems remains unexplored, particularly for spoofed voices of Urdu and Hindi speakers, underscoring the need to develop such spoofed voice datasets. Given that India and Pakistan are among the world's largest diaspora countries, with India leading globally in 2024, this further underscores the need to develop multilingual code-switching datasets for Urdu and Hindi speakers.

This paper introduces a multilingual spoofed speech (MSS) dataset featuring code-switching between Urdu, Hindi, and English for spoof detection. The dataset contains 432,215 spoofed and 40,271 bona fide utterances from 154 speakers. The spoofed samples of the MSS dataset are generated using the latest VC approaches. The audio samples feature conversational styles specifically tailored for Urdu and Hindi speakers, exhibiting linguistic diversity within a single audio sample, as well as overall diversity in speaking style, pronunciation, background, and environmental noises. An end-to-end

First Author (Muhammad Hamza) and Second Author (Hafsa Ilyas) contributed equally to this work. This work is supported by the UK Engineering and Physical Sciences Research Council [Grant number: EP/Y002288/1].

TABLE I
COMPARISON OF THE PROPOSED MSS DATASET WITH OTHER VOICE SPOOFING DATASETS.

Dataset	Year	Language	Spoofed Methods	Real	Fake	M	F
ASVspoof-2019 LA [3]	2019	Eng.	TTS=13, VC=06	12,483	109,816	46	61
ASVspoof-2021 LA [4]	2021	Eng.	TTS=13, VC=06	14,816	133,360	30	37
FMFCC-a [5]	2021	Chinese	TTS=11, VC=02	10,000	40,000	58	73
WaveFake [6]	2021	Eng., Jap.	TTS=07	0	117,985	0	2
HABLA [7]	2022	Spanish	TTS=03, VC=03	22,816	58,000	78	84
TIMIT-TTS [8]	2023	Eng.	TTS=12	0	5,160	46	
CFAD [9]	2024	Chinese	TTS=12	38,600	77,200	1212	
MLAAD [10]	2024	38 Lang.	TTS=82	0	154,000	—	
ASVspoof-5 [11]	2024	Eng.	TTS=19, VC=06, AT=07	171,901	815,262	964	958
MSS (Ours)	**2025**	**Eng., Urdu, Hindi**	**VC=05**	**40,271**	**433,000**	**99**	**55**

baseline framework, named WavSpeech-AASIST, integrating self-supervised learning (SSL) models as the front end in the AASIST model, is also presented to detect voice spoofing attacks. The main contributions of this research work are as follows:

1) A multilingual dataset of spoofed voices is presented in Urdu, Hindi, and English, featuring the conversational styles of bilingual Urdu and Hindi speakers.
2) A robust WavSpeech-AASIST framework is proposed, incorporating pre-trained wav2vec and UniSpeech SSL models in the AASIST framework to detect multilingual code-switching spoofed utterances effectively.
3) A comparative analysis of existing spoofing datasets and contemporary methods is conducted to demonstrate the MSS dataset's distinctiveness and the effectiveness of our WavSpeech-AASIST framework.

The rest of the paper is organized as follows: Section II introduces the MSS dataset. Section III presents the proposed baseline method. Experiments and results are discussed in Section IV. Finally, Section V provides the conclusion.

II. MULTILINGUAL SPOOFED SPEECH DATASET

The proposed multilingual spoofed speech dataset (MSS) comprises bona fide and spoofed utterances in Urdu, Hindi, and English. The bona fide audio samples are sourced from the publicly available MAV-Celeb dataset (licensed under the MIT license) [13], which features utterances from 154 celebrities (extracted from YouTube videos) with diverse background noises and challenging language alternations. To generate realistic spoofed audio samples that preserve the multilingual code-switching and multispeaker attributes of the bona fide samples, five different VC models are employed. A brief description of each method is provided as follows:

Free-VC [14] is a one-shot VC approach using the VITS framework for waveform reconstruction, WavLM features, and a bottleneck extractor for content extraction. It enhances the content information through spectrogram-resized-based data augmentation, resulting in improved robustness.

Diff-Hier-VC [15] is a hierarchical VC method using diffusion models DiffPitch and DiffVoice. DiffPitch generates the target voice's pitch, while DiffVoice creates the Mel-spectrogram based on the pitch. The converted speech is generated using the target voice style, enhanced through a denoising process.

HierSpeech++ [16] is a zero-shot SS method comprised of semantic modeling, a hierarchical speech synthesizer, and speech super-resolution (SpeechSR). The semantic model extracts the semantic representation and pitch information, which the synthesizer uses to generate the speech in the target voice style. SpeechSR is used to create high-resolution synthetic speech.

KNN-VC [17] KNN-VC is a simple and robust VC model based on k-nearest neighbors regression. It extracts the self-supervised embeddings of source and target utterances, replaces the source embedding's frame with its nearest neighbor from the target, and then uses a vocoder to generate converted speech.

Seed-VC [18] Seed-VC is a zero-shot VC approach that uses diffusion transformers for in-context learning and capturing target speech attributes. It employs an external timber shifter to perturb the source speech timber, resulting in natural-sounding synthetic speech with high speaker similarity.

A. Dataset Statistics / Overall Statistics

For the MSS dataset, the audio samples from the MAV-Celeb dataset [13] were divided into training, development, and evaluation sets with a 25:25:50 split ratio, respectively. The split is consistent with the design philosophy of the ASVspoof challenges ([3], [4], [11], [12]), where the evaluation partition is deliberately much larger than training and

TABLE II
STATISTICS OF THE MSS DATASET, IN TERMS OF THE NUMBER OF UTTERANCES AND SPEAKERS.

Dataset	Split	Bona fide	Spoofed	Male	Female
Urdu	Train	3,418	15,300	11	7
	Dev	3,857	13,600	11	6
	Eval	5,431	59,500	21	14
Hindi	Train	2,287	20,500	14	7
	Dev	2,520	21,000	14	7
	Eval	3,329	85,895	28	14
English	Train	5,126	36,300	25	14
	Dev	5,625	34,520	25	13
	Eval	8,678	145,600	49	28

development. This ensures reliable benchmarking and encourages models to generalize rather than overfit. The number of male and female speakers in each subset is distributed accordingly to exclude the bias. We generated ten spoofed audios per speaker using the five selected VC models. For each speaker, one audio sample was designated as the target audio, while 10 randomly selected audio samples from the other speakers were used as source audio, one at a time, to generate the spoofed samples. The overall statistics of the proposed MSS dataset are presented in Table II. The MSS dataset includes 154 speakers (70 Urdu and 84 Hindi). The English subset comprises speakers of both Urdu and Hindi. The audio samples are 101,106 for Urdu, 135,531 for Hindi, and 235,849 for English. All audio samples are in FLAC format, with an average length of 15 seconds, a varied bit rate, and a sampling rate of 16 kHz.

B. Dataset Analysis

A comparative analysis of the MSS dataset with the existing spoofing datasets in Table I reveals that most datasets are monolingual, focusing only on English. Though the MLAAD dataset includes 38 languages, it does not include Urdu, and spoofed samples are generated using only TTS methods. In contrast, the MSS dataset is multilingual, including Urdu, Hindi, and English. More importantly, the existing datasets only account for the accents of native English speakers. In contrast, the MSS dataset specifically includes the English-speaking styles of Urdu and Hindi speakers, highlighting the uniqueness of the English subset within the MSS dataset. Urdu and Hindi speakers have diverse English-speaking styles, frequently use code-switching between their native languages and English, and pronunciation varies across individuals. We did not use TTS models to generate spoofed audios in our dataset, as TTS approaches are designed to synthesize speech in a single language, lacking the ability to mimic the natural conversational dynamics of bilingual Urdu and Hindi speakers. TTS methods cannot effectively preserve natural voice prosodic information, multilingual conversational styles, and sound robotic. Due to these facts, we only utilized VC models to generate the spoofed audios in our MSS dataset. The comparative analysis in Table I highlights the distinctiveness and importance of our MSS dataset among the other existing spoofing datasets.

C. Dataset Evaluation

The proposed MSS dataset was evaluated in different aspects. We first estimated the frequency of code-switching and signal-to-noise ratio (SNR) in each set. Toward this, we randomly selected subsets of audio samples from each set to analyze the frequency of code-switched samples in the Urdu, Hindi, and English sets of the MSS dataset. We marked utterances as code-switched if there exist 15-20% code-switched phrases. Our analysis reveals that approximately 50% of audio samples are code-switched in each set, thereby emphasizing the significant presence of code-switching across Urdu, Hindi, and English sets. Additionally, it is also observed that the audio utterances in the Hindi set demonstrate a higher frequency of language switching within an audio sample compared to the audio samples in the Urdu set of the MSS dataset. Secondly, to analyze the audio quality, we estimated the SNR for each set of the MSS dataset using a voice activity detection (VAD) based energy ratio method. This analysis reveals that 26% of utterances in the Urdu set, 31.59% of utterances in the English set, and 37.87% of the utterances in the Hindi set have an SNR of less than 20dB. This signifies that the Urdu set has the highest proportion of clean audio utterances. However, the Hindi set has a higher proportion of noisy samples.

Next, the spectral composition of audio samples in each set was visualized to highlight the differences and similarities between the bona fide and spoofed audio samples in the MSS dataset. For this, the Mel frequency cepstral coefficient (MFCCs) of the audio samples are extracted and visualized through t-SNE (t-distributed stochastic neighbor embedding). MFCCs are widely used in audio signal processing to capture the power spectrum of audio signals in a manner that aligns with human hearing. We reduced the dimensions of MFCCs by setting the perplexity value to 40, and then plotted the reduced dimensions using the t-SNE method. Fig. 1 - 3 present the 3D t-SNE visualization of the MFCC features for bona fide and spoofed utterances in the Urdu, Hindi, and English sets of the MSS dataset, respectively. It can be observed that substantial overlap exists between bona fide and spoof samples in each set of the MSS dataset, highlighting the resemblance between spoof audios and bona fide ones. Furthermore, Seed-VC, HierSpeech++, and Free-VC generated spoofed samples are closely related to the bona fide ones across all sets of the MSS dataset. This analysis highlights that the Seed-VC and HierSpeech++ spoofing attacks in the MSS dataset are the most difficult to detect, as these attacks generate spoofed audios with characteristics that most closely resemble those of bona fide audios. Overall, t-SNE visualizations indicate the realism of spoofed audio signals, which can pose substantial challenges to spoofing countermeasures.

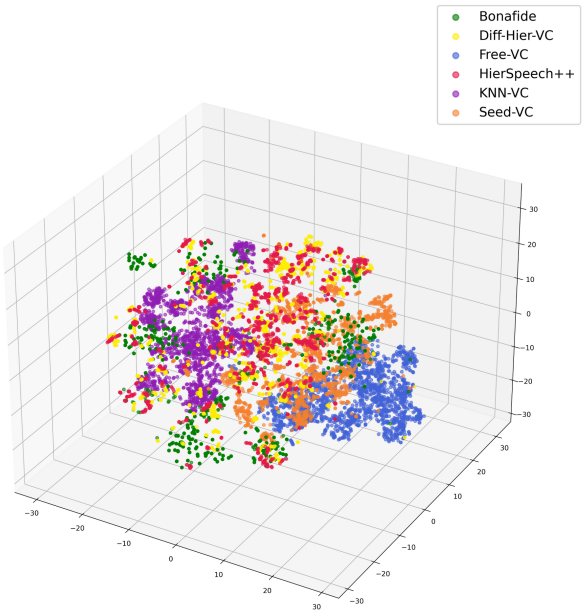

Fig. 1. t-SNE visualization of bona fide and spoof samples of the Urdu set of the MSS dataset.

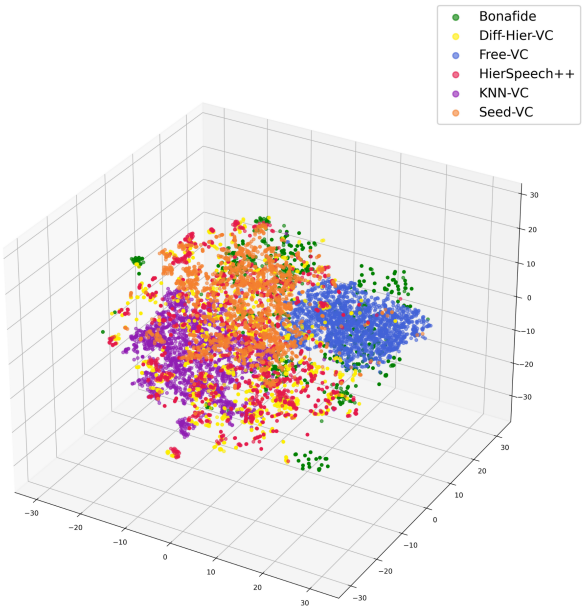

Fig. 3. t-SNE visualization of bona fide and spoof samples of the English set of the MSS dataset.

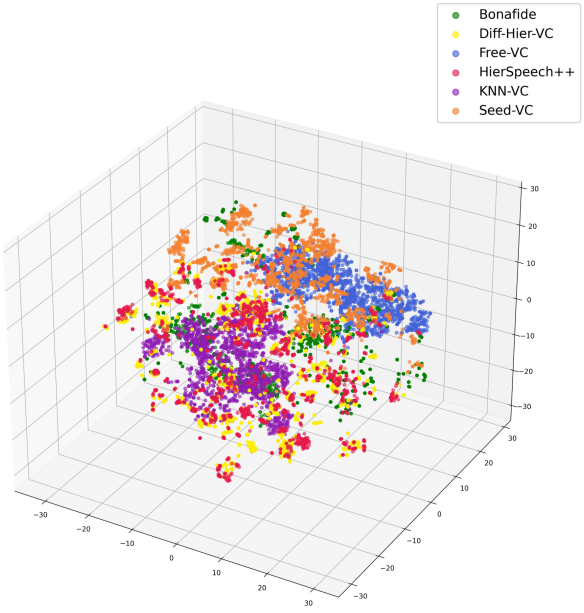

Fig. 2. t-SNE visualization of bona fide and spoof samples of the Hindi set of the MSS dataset.

III. Proposed Methodology / Baseline

This work presents a multi-embedding SSL-based spoofing detection framework, termed WavSpeech-AASIST, as shown in Fig. 4. The framework consists of an SSL feature extractor, a residual encoder, a self-attention aggregation layer, and a spectro-temporal graph attention network, each of which is described in the following subsections.

A. SSL Feature Extractor

To extract the SSL representations from the raw audio I, we introduced the SSL model-based frontend in the AASIST framework. Pre-trained wav2vec-xlsr [19] and UniSpeech-sat [20] models are used to extract complementary representations from the audio signals. The extracted representations from the SSL models are combined to form a unified feature vector $f_{wavSpeech}$, encompassing the linguistic, phonetic, temporal, and speaker-aware representations essential to accurately detect spoofed speech. Mathematically, it can be represented as:

$$f_{wavSpeech} = f_{wav} || f_{speech} \tag{1}$$

where f_{wav} and f_{speech} refer to the features from the wav2vec and UniSpeech models, respectively. The unified feature map $f_{wavSpeech}$ is then processed through a post-processing block comprising a layer for adding channel dimension, max pooling, batch normalization (BN), and SeLU activation layers. The post-processing block transforms the feature map to spectro-temporal representations $R_{wavSpeech}$, fed to the residual encoder.

B. Residual Encoder and Self-Attention Aggregation Layer

A residual encoder comprising six residual blocks is employed in the proposed WavSpeech-AASIST to extract high-level representations $F_{wavSpeech} \in \mathbb{R}^{c \times s \times t}$, where the number of channels, spectral bins, and time-frequency are represented as c, s, and t, respectively. A weighted attention matrix AM_w is generated by implementing a self-attention-based aggregation layer to process the representations $F_{wavSpeech}$ through a Conv layer followed by SeLU, BN, a Conv layer, and a Softmax function. Finally, the weighted sum of representations

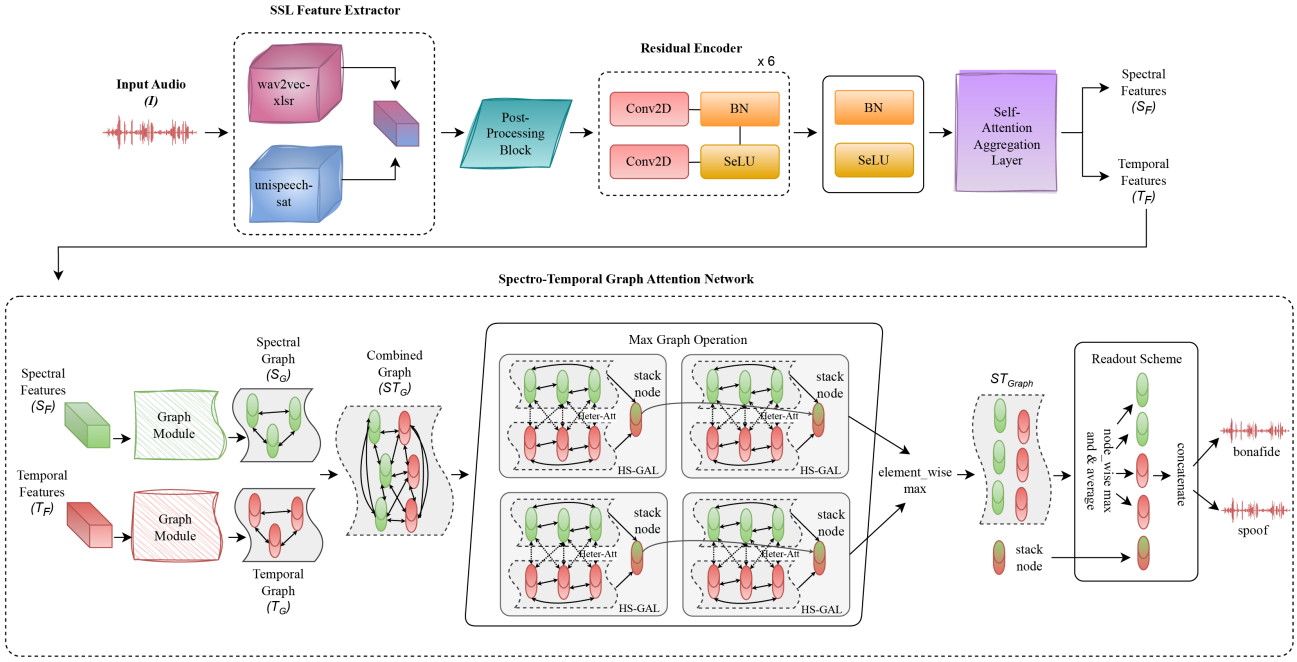

Fig. 4. Architecture of the proposed WavSpeech-AASIST spoofing detection framework.

$F_{wavSpeech}$ is computed using the element-wise multiplication with AM_w, across time t and spectral s domain, to construct spectral S_F and temporal T_F representations, respectively.

C. Spectro-Temporal Graph Attention Network

The spectro-temporal graph attention network consists of spectral and temporal graph modules, max graph operation (MGO), and a readout scheme. Each graph module includes a graph attention network and a graph pooling (GP) layer. The obtained S_F and T_F are processed through corresponding graph modules to construct spectral S_G and temporal T_G graphs, which are combined to build a heterogeneous spectro-temporal ST_G graph. MGO processes ST_G with two parallel branches of heterogeneous stacking graph attention layers (HS-GALs) followed by the GP layer. Each HS-GAL includes a heterogeneous attention mechanism to accommodate graph heterogeneity. However, HS-GALs in each branch of MGO shared the common stack node to preserve the information in S_G and T_G. Then, the elementwise maximum operation is applied to generate another heterogeneous spectro-temporal graph ST_{Graph}. This ST_{Graph} is processed through a readout scheme to extract four nodes by applying node-wise maximum and average operations to the nodes belonging to S_G and T_G, which are finally concatenated with a stack node. The final prediction (bona fide and spoofed) is generated through the last hidden, fully connected layer.

IV. Experiments and Results

This section presents the experiment conducted using the proposed MSS dataset to evaluate the performance of our WavSpeech-AASIST method. We considered the standard evaluation measure equal error rate (EER) [4] utilized in the audio spoofing detection domain for performance evaluation. EER represents the network's ability to distinguish between spoofed and bona fide audio as the point where the false acceptance and rejection rates are equal.

A. Experimental Setup

The model was trained on two classes (spoofed and bona fide) with a weighted categorical cross-entropy loss and a batch size of 8. The training hyperparameters are as follows: optimizer: Adam, learning rate: 0.000001, and weight decay: 0.0001. The model was trained for 25 epochs and the checkpoint with the lowest validation loss was selected for evaluation. All experiments were conducted on an RTX 3090 GPU machine.

B. Performance Evaluation

To evaluate the performance of the proposed WavSpeech-AASIST framework, we conducted two experiments utilizing the proposed MSS and ASVspoof-2019 LA datasets.

The first experiment is conducted on the Urdu, Hindi, and English sets of the MSS dataset. The model is trained on the training and development subsets of the Urdu set and then evaluated on the Urdu evaluation subset. Likewise, we trained and evaluated the model on the Hindi and English sets of the MSS dataset. The results in terms of EER are reported in Table IV. The results demonstrate that the proposed model can effectively distinguish between bona fide and spoofed samples from both Urdu and English sets. However, for the Hindi set of the MSS dataset, WavSpeech-ASSIST achieved the highest EER of 5.55%, indicating that the proposed model struggles to

TABLE III
COMPARATIVE STUDY UTILIZING ASVSPOOF-2019 LA AND OUR PROPOSED MSS DATASET. THE RESULTS ARE REPORTED IN TERMS OF EER (%).

Methods	Urdu	Hindi	English	ASVspoof-2019 LA
RawNet2	5.47	12.28	6.69	5.64
AASIST	3.44	6.89	3.55	0.83
Wav2vec-AASIST	0.40	4.18	1.70	0.47
WavSpeech-AASIST (Proposed)	0.43	5.55	1.39	0.38

TABLE IV
PERFORMANCE OF PROPOSED WAVSPEECH-AASIST ON THE MSS DATASET (EER IN %).

MSS Dataset	Urdu	Hindi	English
EER (%)	0.43	5.55	1.39

distinguish between bona fide and spoofed Hindi utterances. This highlights that the Hindi set is more challenging than the Urdu and English sets of the proposed MSS dataset. In Hindi audio samples, the speakers frequently shift between English and their native language; however, this switching is infrequent in Urdu samples. The language switch can cause a prosodic and phonetic shift, which can be difficult for the model to learn. Moreover, unlike the Urdu set, most bona fide utterances in the Hindi set contain background music and overlapping voices, resulting in noise induction in the generated spoofed samples. Such attributes of the audio samples in the Hindi set of the MSS dataset enhance complexity, potentially hindering the accurate detection of spoofed and bona fide samples in the Hindi set.

In the second experiment, the ASSIST framework is trained on the training and development sets of the ASVspoof-2019 LA dataset and then evaluated on the evaluation set. The proposed Wavspeech-AASIST achieves an EER of 0.38% on the ASVspoof-2019 LA dataset, highlighting the effectiveness of the proposed model in detecting spoofed audio on the standard audio spoofing dataset.

Notably, the EER of the proposed WavSpeech-AASIST on the MSS dataset is higher than that of the ASVspoof-2019 LA dataset, likely due to the code-switching aspect of the MSS dataset. Furthermore, ASVspoof audio samples, being monolingual utterances, also incorporate silent portions, causing improved performance on ASVspoof datasets. Overall, the results indicate that code-switching in the MSS dataset increases the complexity of the audio samples and has an impact on the performance of detecting spoofed utterances.

C. Comparative Study

A comparative analysis of the proposed WavSpeech-AASIST against other baseline methods [21]–[23] is conducted using the ASVspoof-2019 LA and our proposed MSS dataset. Particularly, the other approaches are RawNet2 [21], AASIST [22] and wav2vec-AASIST [23]. To contrast and compare the performance, the baseline methods are trained on the ASVspoof-2019 LA and MSS dataset using the same experimental protocols as mentioned in Section IV-B. The results in terms of EER are presented in Table III.

RawNet2 exhibits the weakest overall performance across both datasets. In contrast, the proposed WavSpeech-AASIST, which fuses wav2vec and UniSpeech embeddings, achieves superior results on the ASVspoof-2019 LA dataset and the English subset of the MSS dataset. Interestingly, wav2vec-AASIST obtains lower EERs on the Urdu and Hindi subsets, suggesting language-specific sensitivities in SSL representations. Despite integrating two SSL feature streams, WavSpeech-AASIST performs slightly worse than wav2vec-AASIST on these subsets. We attribute this to representational redundancy and potential interference between wav2vec and UniSpeech features, as well as the possibility that wav2vec alone encodes spoof-relevant cues more effectively for Urdu and Hindi. This observation underscores that straightforward concatenation of embeddings does not guarantee consistent gains and highlights the need for more adaptive or attention-based fusion strategies. Overall, the comparative study confirms that AASIST-based architectures consistently surpass RawNet2, while the use of SSL features substantially enhances spoofing detection performance. Notably, WavSpeech-AASIST demonstrates the promise of multi-embedding SSL features for robust spoof detection, even as the Hindi subset remains the most challenging within the MSS dataset.

V. CONCLUSION

This work introduces the MSS dataset, comprising spoofed Urdu, Hindi, and English utterances featuring the speaking accents and bilingual conversational style of Hindi and Urdu speakers. The dataset captures multi-speaker scenarios, linguistic diversity in a single audio utterance, varied speaking styles, and background noises. In addition, we propose WavSpeech-AASIST, a robust framework that leverages multi-embedding SSL features for spoofing detection. Experimental findings highlight the challenging nature of the Hindi set of the proposed MSS dataset. Future work will focus on enriching the dataset by including replay and adversarial attacks.

REFERENCES

[1] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A survey on speech deepfake detection," *ACM Computing Surveys*, vol. 57, no. 7, pp. 1–38, 2025.

[2] A. Triantafyllopoulos, B. W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E. André *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1355–1381, 2023.

[3] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[4] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[5] Z. Zhang, Y. Gu, X. Yi, and X. Zhao, "FMFCC-a: A challenging Mandarin dataset for synthetic speech detection," in *International Workshop on Digital Watermarking*. Springer, 2021, pp. 117–131.

[6] J. Frank and L. Schönherr, "Wavefake: A data set to facilitate audio deepfake detection," *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[7] P. T. Flórez, R. Manrique, and B. P. Nunes, "HABLA: A dataset of Latin American Spanish accents for voice anti-spoofing," in *Proc. Interspeech*, 2023, pp. 1963–1967.

[8] D. Salvi, B. Hosler, P. Bestagini, M. C. Stamm, and S. Tubaro, "TIMIT-TTS: A text-to-speech dataset for multimodal synthetic media detection," *IEEE Access*, vol. 11, pp. 50 851–50 866, 2023.

[9] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, and R. Fu, "CFAD: A Chinese dataset for fake audio detection," *Speech Communication*, vol. 164, p. 103122, 2024.

[10] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "MLAAD: The multi-language audio anti-spoofing dataset," *arXiv preprint arXiv:2401.09512*, 2024.

[11] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen *et al.*, "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," *arXiv preprint arXiv:2408.08739*, 2024.

[12] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 15, p. 3750, 2014.

[13] M. S. Saeed, S. Nawaz, M. Moscati, R. K. Das, M. S. Tahir, M. Z. Zaheer, M. I. Liaqat, M. H. Khan, K. Nandakumar, M. H. Yousaf, and M. Schedl, "A synopsis of fame 2024 challenge: Associating faces with voices in multilingual environments," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 11333–11334. [Online]. Available: https://doi.org/10.1145/3664647.3688978

[14] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "Diff-HierVC: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation," *International Speech Communication Association*, pp. 2283–2287, 2023.

[16] S.-H. Lee, H.-Y. Choi, S.-B. Kim, and S.-W. Lee, "Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[17] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," *Proc. Interspeech*, pp. 2053–2057, 2023.

[18] S. Liu, "Zero-shot voice conversion with diffusion transformers," *arXiv preprint arXiv:2411.09943*, 2024.

[19] 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.

[20] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6152–6156.

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

[22] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

[23] 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," *arXiv preprint arXiv:2202.12233*, 2022.