The Vietnamese Spoofing-aware Speaker Verification Challenge 2025: Summary and Results

Phuong Tuan Dat, Hoang Long Vu, Nguyen Thi Thu Trang[†]
SoICT, Hanoi University of Science and Technology
{phuongtuandat2915, longvu200502}@gmail.com, trangntt@soict.hust.edu.vn

Abstract

The VLSP 2025 Vietnamese Spoofing-Aware Speaker Verification (VSASV) Challenge extends the study of spoofing-aware speaker verification (SASV) to Vietnamese, a low-resource language with limited anti-spoofing data resources. Building upon prior SASV challenges, VSASV introduces an evaluation framework encompassing both bonafide and spoofed trials, including replay, voice conversion (VC), text-to-speech (TTS), and adversarial attacks. The dataset design simulates realistic low-resource conditions where only replay attacks are available during training, while test sets include previously unseen spoofing types to assess model generalization capabilities. The challenge provides open-source baselines integrating ECAPA-TDNN for Automatic Speaker Verification (ASV) and XLSR-Conformer+TCM for Countermeasures (CM), fused at the score level. Results from 9 registered teams demonstrate that the majority of submitted systems outperform the baselines, confirming the effectiveness of fusion-based and Self-supervised learning (SSL) driven approaches. VSASV encourages continued research on spoofing-aware speaker verification and deepfake detection for Vietnamese and other low-resource languages, paving the way toward more reliable and inclusive speech authentication technologies.

1 Introduction

ASV (Saquib et al., 2010) systems aim to determine whether a given speech segment was produced by a claimed speaker. As one of the most convenient, natural, and non-intrusive biometric modalities, ASV has gained widespread adoption in numerous applications, particularly in telephony-based authentication and access control systems (Anjum and Swamy, 2017). Despite their effectiveness in distinguishing between target and impostor trials, ASV systems remain vulnerable to spoofed utterances—speech signals that have been manipulated, synthesized, or generated using advanced VC (Azzuni and Saddik, 2025) or TTS (Tan et al., 2021) techniques. Such spoofing attacks pose critical threats to the reliability and security of ASV systems. Research into spoofing countermeasures (CMs) has advanced significantly over the past decade, largely driven by the ASVspoof (Wang et al., 2020) (Yamagishi et al., 2021) (Wang et al., 2024) initiative and its associated challenge series, which have established benchmarks for assessing spoofing detection performance. Traditionally, CMs are implemented as standalone binary classifiers designed to differentiate between bonafide and spoofed utterances, and are combined with ASV systems in a cascaded or gated manner. While such integration can enhance robustness against spoofing attacks, it also introduces trade-offs: overly strict countermeasures may reject genuine target trials, thereby degrading overall system usability. To address this challenge, the SASV (weon Jung et al., 2022b) paradigm advocates for a joint evaluation and optimization framework in which ASV and CM components are developed in tandem. This integrated approach acknowledges the interdependence between subsystems and aims to achieve reliable ASV performance under both spoofed and bonafide conditions. The SASV Challenge has thus encouraged two complementary research directions: (1) fusion-based systems, which combine existing ASV and CM models through learned fusion strategies, and (2) single integrated architectures, which jointly learn speaker identity and spoofing awareness within a unified latent representation. Building upon the foundations laid by SASV for English corpora, the VLSP 2025 Challenge on VSASV extends this line of research to Vietnamese, a low-resource language with distinct phonetic and prosodic characteristics. Compared to high-resource languages such as English, Viet-

[†]Corresponding author.

namese presents unique challenges for both ASV and CM development, including limited availability of annotated spoofing data, language-specific acoustic variability, and tonal complexity. Consequently, the VSASV Challenge represents an important step toward inclusive spoofing-aware verification research. By focusing on Vietnamese, this challenge aims to foster innovation in developing robust, data-efficient, and spoofing-aware speaker verification solutions for tonal and low-resource languages. The ultimate goal is to promote research that enhances the reliability, security, and fairness of ASV systems in resource-constrained settings.

2 Datasets and evaluation metrics

2.1 Datasets

The training and testing datasets for the VSASV Challenge are constructed based on a combination of publicly available Vietnamese speech corpora, including Vietnam-Celeb (Pham et al., 2023), VoxVietnam (Vu et al., 2025), and VSASV (Hoang et al., 2024). These datasets collectively provide a diverse and representative collection of Vietnamese speech samples that cover variations in speakers, accents, and recording conditions. The inclusion of both bonafide and spoofed utterances enables participants to develop, train, and evaluate spoofing-aware speaker verification systems that are robust under real-world conditions. In addition, we augment the dataset with spoofed data generated through diverse spoofing techniques, including replay attacks, VC, and TTS synthesis, to ensure broader coverage of potential spoofing scenarios. Table 1 presents the overall statistics of the VSASV Challenge dataset. The Public Test set comprises 491,676 audio pairs, while the Private Test set contains 557,516 audio pairs. The speakers in the training set and the two test sets are mutually exclusive, ensuring no speaker overlap across subsets.

Subset	Utterance type	# Utterances	# Hours
Train	Bonafide	71,617	152.89
	Spoof	29,750	50.38
Public Test	Bonafide	235,577	353.29
	Spoof	256,099	452.76
Private Test	Bonafide	185,577	276.42
	Spoof	371,939	503.24

Table 1: Statistics of the VSASV dataset across subsets and utterance types.

Table 2 presents the distribution of spoofing utterances across different spoof types and dataset partitions. The design of the VSASV dataset simulates a realistic low-resource scenario where the availability of spoofed data is limited, and only a single spoofing type—replay attacks—is present in the training set. This setup encourages participants to develop models capable of generalizing beyond seen spoof types. In contrast, the test sets introduce additional out-of-domain spoofing techniques, including VC, TTS, and adversarial attacks, which are unseen during training. The Public Test subset emphasizes replay attacks, mirroring the spoofing condition observed in training, while the Private Test subset contains a higher proportion of unseen spoofing types. This design allows for a comprehensive evaluation of system robustness and generalization capability under mismatched spoofing conditions.

Subset	VC	Replay	Adversarial	TTS
Train	0	29,750	0	0
Public Test	55,099	150,000	50,000	1,000
Private Test	210,139	5,000	150,000	6,800

Table 2: The number of utterances of each spoof type.

Figure 1 illustrates the data generation pipeline for the VSASV Challenge dataset. The process begins with bonafide audio samples extracted from public Vietnamese speaker verification datasets. These samples are first processed through a speaker embedding module utilizing the ECAPA-TDNN model to compute pairwise cosine similarity scores between speakers. Based on these similarity scores, speakers are partitioned into two groups: the top 30% of speaker pairs with the highest similarity scores, which are allocated to the test sets, and the remaining 70%, which comprise the training set. For the high-similarity group in the test sets, spoofed utterances are generated using advanced synthesis techniques including VC, TTS, and adversarial attacks, in addition to replay attacks. This design choice ensures that the most challenging spoofing scenarios - where target and source speakers share acoustic similarities - are adequately represented in the evaluation phase. For the training set comprising the remaining 70% of speakers, only replay attacks are applied. This stratified approach to spoof generation enables a balanced evaluation of system robustness across varying levels of speaker similarity, while maintaining realistic proportions of different spoofing types and simulating

low-resource training conditions.

2.2 Evaluation metrics

The performance of systems participating in the VSASV Challenge is primarily evaluated using the Equal Error Rate (EER), denoted as SASV-EER, which serves as the main metric. Following previous SASV studies, the task is formulated as a binary classification problem: distinguishing target trials from non-target ones. The non-target category comprises both bonafide impostor trials and spoofed non-target trials, each contributing to potential increases in the false acceptance rate when countermeasures are insufficient. In addition to the SASV-EER, two supplementary metrics are reported to provide a more detailed assessment of system performance. The Speaker Verification EER (SV-EER) measures errors arising solely from target and bonafide non-target trials, thereby reflecting performance in the absence of spoofing attacks. Conversely, the Spoofing EER (SPF-EER) is computed using target and spoofed non-target trials, capturing system robustness under spoofing conditions. These complementary metrics allow for a deeper understanding of system reliability across different scenarios—one dominated by genuine impostor trials and the other by spoofed inputs.

3 Baseline Model

3.1 ASV sub-system

The ASV baseline module employed in the VSASV Challenge is built upon the ECAPA-TDNN (Desplanques et al., 2020) architecture, a state-of-the-art speaker embedding extractor known for its efficiency and robustness. The model leverages a Res2Net-based (Gao et al., 2021) backbone integrated with Squeeze-and-Excitation (SE) (Hu et al., 2019) blocks, enabling effective channel-wise feature recalibration and improved representational power. To capture both global and contextual information, a channel- and context-dependent statistics pooling layer is utilized, followed by multi-layer feature aggregation that transforms frame-level representations into comprehensive utterance-level embeddings. The input features are 80-dimensional Mel-filterbank coefficients, which have been widely adopted in speaker verification tasks. After pooling, speaker embeddings are projected through an affine transformation using a fully connected layer to form the final ASV representations. To enhance generalization and robustness against acoustic variability, data augmentation techniques are applied using room impulse responses (RIRs) (Arellano et al., 2025) and additive noise samples drawn from the MUSAN (Snyder et al., 2015) corpus. The training follows standard open-source recipes from previous ECAPA-TDNN implementations[1], ensuring reproducibility and comparability. This baseline provides a strong foundation for evaluating spoofing-aware speaker verification models within the VSASV framework.

3.2 CM sub-system

In this study, we adopt the XLSR-Conformer+TCM model as our baseline architecture, following the design introduced in (Truong et al., 2024). The model builds upon a cross-lingual self-supervised representation (XLS-R) backbone and a Conformer encoder to effectively capture both temporal and contextual information from speech signals. Specifically, the pre-trained XLS-R model, derived from wav2vec 2.0 (Baevski et al., 2020), serves as a feature extractor to generate rich, contextualized speech representations. Given an input waveform (O), the SSL module produces a sequence of feature vectors $X = \mathrm{SSL}(O) = x_t \in \mathbb{R}^D \mid t = 1, \ldots, T$ where D denotes the embedding dimension.

To adapt these high-dimensional features for downstream processing, they are projected to a lower-dimensional space using a linear transformation followed by a SeLU activation, forming

$$\tilde{X} = \mathrm{SeLU}(\mathrm{Linear}(X))$$

The transformed features are then fed into a stack of Conformer blocks, where each block integrates multi-head self-attention (MHSA) and convolutional modules to model both global and local dependencies. For classification purposes, a learnable class token is prepended to the feature sequence, enabling the model to aggregate global information across time steps. The output state of this class token after the final Conformer layer is subsequently passed through a linear classifier to predict whether the input is bonafide or spoofed.

A key enhancement of this baseline lies in the Temporal-Channel Modeling (TCM) mechanism, which fuses channel-level and temporal representations directly within the MHSA module. By incorporating this cross-dimensional interaction, the XLSR-Conformer+TCM is able to jointly model

[1] https://github.com/TaoRuijie/ECAPA-TDNN.git

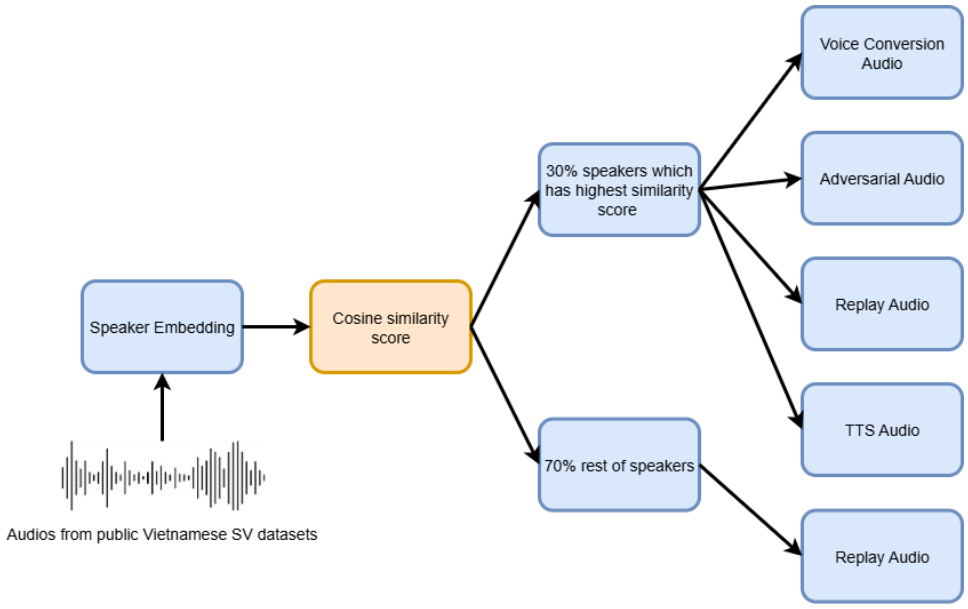

Figure 1: The pipeline of building VSASV challenge dataset.

temporal dynamics and inter-channel dependencies, leading to significant improvements in detecting synthetic speech. As demonstrated in (Truong et al., 2024), this architectural modification yields a relative error rate reduction of 26% on the ASVspoof 2021 logical access benchmark, showcasing its effectiveness in anti-spoofing scenarios. This CM sub-system can be reproduced via open-source link[2].

3.3 Baseline system

The SASV 2022 Challenge (weon Jung et al., 2022b) introduced two open-source baseline systems: (i) score-sum fusion and (ii) DNN back-end fusion. Both approaches employ independently trained ASV and CM sub-systems. The score-sum fusion baseline operates at the score level, where the outputs from the ASV and CM modules are combined through simple score addition. This fusion strategy is parameter-free, requiring no additional training or fine-tuning. Specifically, ASV scores are derived from the cosine similarity between enrollment and test utterances, while CM scores may optionally be passed through a softmax function to normalize their range to (0, 1). In this work, we adopt the score-sum fusion method as our baseline system.

[2]https://github.com/ductuantruong/tcm_add.git

4 Challenge Results

Out of a total of 30 registered teams, we received 9 valid submissions for the Public Test set and 8 submissions for the Private Test set. Table 3 presents the performance results on the Public Test set, while Table 4 summarizes the corresponding results on the Private Test set.

Table 3: Performance results (EER) on the Public Test set.

#	Team ID	EER (%)
1	SA-SVBK	20.79
2	Brosh	22.04
3	ChatJLPT	24.75
4	SV++	26.91
5	HuevsBentre	27.04
6	NLP Noobs	27.09
7	Doraemon	28.82
8	Arrebol	28.88
9	Baseline	33.88
10	RD	36.85

From a total of 10 valid submissions, almost participating teams achieved better results compared to the baseline system on both the Public and Private Test sets. Specifically, as shown in Table 3, 9 out of 10 systems surpass the baseline performance on the Public Test set, while Table 4 demonstrates that 8 out of 9 teams outperform the baseline on the Private Test set. These results highlight the effectiveness of the proposed approaches, where

Table 4: Performance results (EER) on the Private Test set.

#	Team ID	EER (%)
1	SV++	17.78
2	SA-SVBK	17.86
3	ChatJLPT	24.37
4	Brosh	24.60
5	RD	29.83
6	NLP Noobs	30.63
7	Arrebol	32.48
8	Baseline	36.78
9	TQ	43.65

most systems exhibit strong generalization across evaluation conditions.

Notably, the best-performing systems achieved substantial improvements in terms of EER. On the Public Test set, the top system (trungnt_dsai) obtained an EER of 20.79%, representing a relative reduction of approximately 38.6% compared to the baseline (33.88%). Similarly, on the Private Test set, the leading system (hoanglp123) achieved the lowest EER of 17.78%, indicating a 51.5% relative improvement over the baseline (36.78%).

Overall, the results are highly encouraging and demonstrate the competitiveness of submitted systems. The consistent ranking across both test sets also suggests that these top-performing models exhibit good robustness and generalization when facing unseen spoofing conditions.

5 Solution Strategies

The top two systems in both the Private Test sets, as shown in Tables 3 and 4, share several key design characteristics that contribute to their superior performance. First, both approaches adopt a fusion framework that combines independent ASV and CM sub-systems, allowing the overall system to leverage the complementary strengths of speaker verification and spoofing detection components. Second, both perform fusion at the score level, which has proven to be an effective and lightweight strategy for integrating heterogeneous subsystem outputs without the need for additional training or complex parameter tuning. Third, both systems employ SSL-based architectures for their CM subsystems, motivated by the recent success of self-supervised learning models in deepfake and spoofing detection tasks (Tak et al., 2022) (Phuong et al., 2025) (Xiao and Das, 2025). The strong representational power of SSL models enables more discriminative feature learning, ultimately enhancing the overall robustness and generalization capability of the fused system.

Despite sharing similar overall design strategies, the top two systems differ in several key aspects that contribute to their unique performance characteristics. First, the top-ranked system employs ERes2NetV2 (Chen et al., 2024) as its ASV subsystem, whereas the second-ranked system adopts the MFA-Conformer (Yang Zhang and Zhiqiang Lv and Haibin Wu and Shanshan Zhang and Pengfei Hu and Zhiyong Wu and Hung-yi Lee and Helen Meng, 2022) architecture.

Second, the two systems also diverge in their data preparation strategies. The top-performing team applies extensive spoofed data augmentation covering three spoofing types - VC, TTS, and Adversarial - to enhance model robustness across diverse attack scenarios. In contrast, the second-ranked system focuses on dataset refinement, employing DBScan-based data cleaning to remove noisy or mislabeled samples, thereby improving training stability and representation quality.

Beyond the top-performing systems, other participating teams explored alternative architectural choices for their fusion frameworks, as shown in Table 5.

#	Team ID	ASV module	CM module
1	SV++	Eres2NetV2	XLSR-Conformer+TCM
2	SA-SVBK	MFA-Conformer	XLSR-Conformer+TCM
3	ChatJLPT	ECAPA	AASIST
4	Brosh	ResNet-48	AASIST
5	RD	ECAPA	AASIST
6	NLP Noobs	RawNet3	AASIST
7	Arrebol	ECAPA	AASIST
8	Baseline	ECAPA	XLSR-Conformer+TCM
9	TQ	ECAPA	XLSR-Conformer+TCM

Table 5: Summary of methodologies of each team in the challenge.

Beyond the top two systems, the remaining teams explored diverse architectural combinations while adhering to the fusion-based framework, as summarized in Table 5. For the ASV module, several teams adopted ECAPA-TDNN, while others experimented with alternative architectures including ResNet-48 (Alexander Alenin and Nikita Torgashov and Anton Okhotnikov and Rostislav Makarov and Ivan Yakovlev, 2022), RawNet3 (weon Jung et al., 2022a), and ERes2NetV2. On the CM side, AASIST emerged as a popular choice among multiple teams, while others leveraged SSL-

based models such as XLSR-Conformer+TCM. Interestingly, several teams employed identical combinations of ASV and CM modules but achieved markedly different performance outcomes. This performance variance among architecturally identical systems suggests that data processing strategies, including augmentation techniques, preprocessing pipelines, training procedures, and hyperparameter configurations, play a more decisive role than model selection alone.

These design variations highlight complementary approaches - one emphasizing data diversity and the other focusing on data quality - both of which effectively contribute to improved spoofing-aware speaker verification performance.

6 Discussion and Conclusion

The VSASV Challenge was established to foster research on SASV for low-resource languages, with a particular focus on Vietnamese. Unlike previous challenges that primarily targeted English, this work aims to address a crucial gap in the field—the scarcity of publicly available spoofing datasets for Vietnamese. This limitation has long hindered the development of robust ASV and countermeasure systems for underrepresented languages. Through the construction of a dedicated Vietnamese dataset that includes both bonafide and spoofed utterances generated via diverse spoofing techniques such as replay, VC, TTS, and adversarial attacks, the VSASV Challenge provides a new testbed for benchmarking and advancing spoofing-aware ASV systems under realistic conditions. The results demonstrate that many participating systems achieved strong performance, substantially outperforming the baseline, thereby validating the feasibility of building reliable ASV and CM systems even under data-limited scenarios. Moreover, the challenge outcomes underscore the synergistic advantage of combining ASV and CM sub-systems, as fusion-based architectures consistently deliver robust performance across both public and private test sets. These findings also suggest that modern SSL-based models, which have shown remarkable success in English deepfake detection tasks, can be effectively adapted to Vietnamese when coupled with appropriate training and augmentation strategies. Looking ahead, further progress will likely depend on expanding the scale and diversity of Vietnamese spoofing datasets. Future efforts should explore jointly optimized, end-to-end architectures that unify ASV and CM into a single framework, thereby reducing reliance on ensemble methods. We hope that the VSASV Challenge will inspire continued research in speaker verification and deepfake detection for low-resource languages, paving the way toward more inclusive and globally reliable speech authentication technologies.

References

Alexander Alenin and Nikita Torgashov and Anton Okhotnikov and Rostislav Makarov and Ivan Yakovlev. 2022. A Subnetwork Approach for Spoofing Aware Speaker Verification. In *Interspeech 2022*, pages 2888–2892.

Z. Khamar Anjum and R. Kumara Swamy. 2017. Spoofing and countermeasures for speaker verification: A review. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 467–471.

Silvia Arellano, Chunghsin Yeh, Gautam Bhattacharya, and Daniel Arteaga. 2025. Room impulse response generation conditioned on acoustic parameters. *Preprint*, arXiv:2507.12136.

Hussam Azzuni and Abdulmotaleb El Saddik. 2025. Voice cloning: Comprehensive survey. *Preprint*, arXiv:2505.00579.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, Shiliang Zhang, and Junjie Li. 2024. ERes2NetV2: Boosting Short-Duration Speaker Verification Performance with Computational Efficiency. In *Interspeech 2024*, pages 3245–3249.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, pages 3830–3834.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. 2021. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662.

Vu Hoang, Viet Thanh Pham, Hoa Nguyen Xuan, Pham Nhi, Phuong Dat, and Thi Thu Trang Nguyen. 2024. VSASV: a Vietnamese Dataset for Spoofing-Aware Speaker Verification. In *Interspeech 2024*, pages 4288–4292.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2019. Squeeze-and-excitation networks. *Preprint*, arXiv:1709.01507.

Viet Thanh Pham, Xuan Thai Hoa Nguyen, Vu Hoang, and Thi Thu Trang Nguyen. 2023. Vietnam-celeb: a large-scale dataset for vietnamese speaker recognition. In *Interspeech 2023*, pages 1918–1922.

Tuan Dat Phuong, Long-Vu Hoang, and Huy Dat Tran. 2025. Pushing the Performance of Synthetic Speech Detection with Kolmogorov-Arnold Networks and Self-Supervised Learning Models. In *Interspeech 2025*, pages 5633–5637.

Zia Saquib, Nirmala Salam, Rekha Nair, Nipun Pandey, and Akanksha Joshi. 2010. A survey on automatic speaker recognition systems. volume 123, pages 134–145.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *Preprint*, arXiv:1510.08484.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *Preprint*, arXiv:2202.12233.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *Preprint*, arXiv:2106.15561.

Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. 2024. Temporal-Channel Modeling in Multi-head Self-Attention for Synthetic Speech Detection. In *Interspeech 2024*, pages 537–541.

Hoang Long Vu, Phuong Tuan Dat, Pham Thao Nhi, Nguyen Song Hao, and Nguyen Thi Thu Trang. 2025. Voxvietnam: a large-scale multi-genre dataset for vietnamese speaker recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi H. Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. 2024. Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, and 21 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech Language*, 64:101114.

Jee weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. 2022a. Pushing the limits of raw waveform speaker recognition. *Preprint*, arXiv:2203.08488.

Jee weon Jung, Hemlata Tak, Hye jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. 2022b. Sasv 2022: The first spoofing-aware speaker verification challenge. *Preprint*, arXiv:2203.14732.

Yang Xiao and Rohan Kumar Das. 2025. Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection. *IEEE Signal Processing Letters*, 32:1276–1280.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 47–54.

Yang Zhang and Zhiqiang Lv and Haibin Wu and Shanshan Zhang and Pengfei Hu and Zhiyong Wu and Hung-yi Lee and Helen Meng. 2022. MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification. In *Interspeech 2022*, pages 306–310.