VNU Journal of Science: Comp. Science & Com. Eng., Vol. 41, No. 2 (2025) 68–85

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

The Vietnamese Spoofing-aware Verification Challenge: A Comprehensive Analysis and Future Work

Phuong Tuan Dat, Hoang Long Vu, Nguyen Thi Thu Trang*

SoICT, Hanoi University of Science and Technology

Received 05th December 2015;
Revised 18th December 2025; Accepted 22nd December 2025

Abstract: The Vietnamese Spoofing-Aware Speaker Verification (VSASV) Challenge series represents the first systematic effort to advance spoof-resistant speaker verification for Vietnamese - a low-resource, highly tonal language characterized by rich phonetic variability. Unlike prior challenges focused on English, VSASV directly addresses the scarcity of publicly available Vietnamese spoofing corpora, a limitation that historically hindered the development of robust automatic speaker verification (ASV) and spoofing countermeasure (CM) systems. Across its 2023 and 2025 editions, VSASV introduces progressively more challenging benchmarks, including multi-corpus bonafide speech, replay attacks, neural voice conversion, modern TTS synthesis, and adversarial perturbations. The 2025 edition further incorporates a speaker-similarity-based partitioning strategy and severe train–test mismatches to emulate realistic attack scenarios. Results from more than 40 participating systems highlight the feasibility of building reliable spoofing-aware ASV pipelines under low-resource conditions, particularly when combining ASV and CM subsystems or leveraging multilingual self-supervised learning (SSL) models. The findings underscore the importance of linguistic properties - especially tonal dynamics - in shaping spoofing vulnerabilities and model generalization. This work provides a comprehensive overview of the VSASV challenge series, synthesizing insights that inform future research on deepfake detection, spoof-robust speech authentication, and inclusive biometric technologies for underrepresented languages.

Keywords: Deepfake Detection, Speaker Verification, Low-resource Languages, Vietnamese Speech Datasets

1. Introduction

Automatic Speaker Verification (ASV) [1] 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

*Corresponding author.
E-mail address: trangntt@soict.hust.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.6492

adoption in numerous applications, particularly in telephone-based authentication and acces control systems [2]. 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 Voice Conversion (VC) [3] or Text-to-Speech (TTS) [4] 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 [5–7] 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 Spoofing-Aware Speaker Verification (SASV) [8] 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 Vietnamese Spoofing-Aware Speaker Verification (VSASV) Challenge series 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, Vietnamese presents unique challenges for both ASV and CM development, including limited availability of annotated spoofing data, language-specific acoustic variability, and tonal complexity inherent to the language's six-tone system.

The VSASV initiative was launched in 2023 as part of the Vietnamese Language and Speech Processing (VLSP) workshop, representing the first comprehensive effort to establish standardized benchmarks for spoofing-aware speaker verification in Vietnamese. The inaugural VSASV 2023 Challenge focused on establishing baseline performance metrics and assessing the applicability of existing ASV and CM methodologies to Vietnamese speech. This initial challenge revealed critical insights into the unique vulnerabilities of Vietnamese ASV systems to spoofing attacks and highlighted the need for language-specific approaches to countermeasure development.

Following the success and lessons learned from VSASV 2023, the VSASV 2025 Challenge builds upon this foundation with enhanced dataset complexity, expanded attack scenarios, and more stringent evaluation protocols. The 2025 iteration introduces several key innovations: (1) a larger and more diverse corpus encompassing multiple recording conditions and channel variations, (2) inclusion of advanced neural vocoder-based synthesis methods reflecting the latest developments in deepfake generation, (3) introduction of adversarial and adaptive attack scenarios designed to challenge state-of-the-art countermeasures, and (4) emphasis on practical deployment considerations such as computational efficiency and real-time processing capabilities.

By examining both VSASV 2023 and 2025 challenges in this work, we aim to provide a

comprehensive overview of the evolution of spoofing-aware speaker verification research for Vietnamese. This retrospective analysis enables us to: (1) trace the progression of system performance across challenge iterations, (2) identify persistent challenges and emerging vulnerabilities specific to low-resource tonal languages, (3) highlight successful methodological innovations and their transferability across languages, and (4) establish future research directions for robust, data-efficient, and spoofing-aware speaker verification solutions in resource-constrained multilingual settings.

The ultimate goal of the VSASV Challenge series is to promote research that enhances the reliability, security, and fairness of ASV systems for underrepresented languages. By fostering innovation in Vietnamese spoofing-aware verification, this initiative contributes to the broader objective of developing inclusive biometric authentication technologies that serve diverse linguistic communities while maintaining high standards of security and usability.

2. VSASV 2023 Dataset

The VSASV 2023 Challenge introduced the first comprehensive benchmark for spoofing-aware speaker verification in Vietnamese. The dataset comprises both training and evaluation sets, incorporating bonafide utterances from real-world scenarios and spoofed samples generated through multiple attack techniques.

2.1. Training Data

The training data consists of two primary components: bonafide utterances for speaker verification model development and spoofed utterances for countermeasure training.

2.1.1. Bonafide Training Data

The bonafide training set was constructed from three Vietnamese speech corpora: Vietnam-Celeb [9], CommonVoice [10], and VIVOS [11]. Vietnam-Celeb, specifically designed for Vietnamese speaker recognition, contains recordings from diverse real-world scenarios including podcasts, interviews, talk shows, and movies, covering all three main Vietnamese dialects (Northern, Central, and Southern). To simulate realistic conditions with potential label noise, utterances were randomly shuffled between speakers in the Vietnam-Celeb subset.

Table 1 presents the detailed statistics of the bonafide training data. The complete bonafide training set encompasses 945 speakers with 147,730 utterances, totaling approximately 225.68 hours of speech.

Table 1. Statistics of bonafide training data in VSASV 2023

Source Corpus	# Speakers	# Utterances	Duration (hours)
Vietnam-Celeb	835	132,424	206.35
CommonVoice + VIVOS	110	15,306	19.33
Total	**945**	**147,730**	**225.68**

2.1.2. Spoofed Training Data

The spoofed training set was generated using two distinct attack methodologies to provide comprehensive coverage of potential spoofing threats. The first category consists of synthesized speech attacks, created by training state-of-the-art voice conversion models based on the Retrieval-based Voice Conversion (RVC) framework[1]. Speaker pairs were constructed from CommonVoice and VIVOS datasets, where utterances from one speaker were transformed to mimic another speaker's voice characteristics.

The second category comprises replay attacks, simulating physical access scenarios where bonafide recordings are played back and re-recorded using various consumer devices such as mobile phones and laptops. This attack type represents a practical threat where attackers may present recorded audio to verification systems.

[1]https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI.git

Table 2 summarizes the spoofed training data statistics, which includes 72 speakers for synthesis attacks (37,292 utterances, 42.84 hours) and 46 speakers for replay attacks (23,320 utterances, 30.03 hours).

Table 2. Statistics of spoofed training data in VSASV 2023

Attack Type	# Speakers	# Utterances	Duration (hours)
Synthesized Speech	72	37,292	42.84
Audio Replay	46	23,320	30.03
Total	**118***	**60,612**	**72.87**

*Number of unique speakers across attack types

2.2. Evaluation Data

The evaluation protocol consisted of two test sets: a public test set for system development and validation, and a private test set for final ranking. Both test sets followed a verification trial format where each sample consists of an enrollment utterance (always bonafide) paired with a test utterance (either bonafide or spoofed).

2.2.1. Trial Configuration

Trials were categorized into three types based on their composition:

- **Positive pairs**: Both enrollment and test utterances are bonafide and originate from the same speaker (Label 1)
- **Bonafide negative pairs**: Both utterances are bonafide but from different speakers (Label 0)
- **Spoofed negative pairs**: Enrollment is bonafide while test utterance is spoofed (Label 0)

A critical distinction between the public and private test sets was the application of adversarial attacks. While 50% of spoofed utterances in the public test set were subjected to gradient-based adversarial perturbations, the private test set employed adversarial attacks on 100% of spoofed samples, significantly increasing the evaluation difficulty.

2.2.2. Test Set Statistics

Table 3 presents the utterance-level statistics, while Table 4 details the trial pair composition for both test sets.

Table 3. Utterance statistics of VSASV 2023 test sets

Utterance Type	Public Test	Private Test
Bonafide	7,256	11,421
Spoofed	5,089	12,102
Total	**12,345**	**23,523**

Table 4. Trial pair statistics of VSASV 2023 test sets

Trial Type	Public Test	Private Test
Positive Pairs	7,587	13,008
Bonafide Negative Pairs	46,140	42,131
Spoofed Negative Pairs	25,393	63,463
Total Trials	**79,120**	**118,602**

Adversarial Attack Rate: 50% (Public), 100% (Private)

The substantial increase in spoofed negative pairs in the private test set (63,463 vs. 25,393) combined with universal adversarial attack application created a particularly challenging evaluation scenario, effectively testing system robustness against both sophisticated synthesis methods and adversarial perturbations. All speakers in the test sets were disjoint from the training set, ensuring proper generalization evaluation.

3. VSASV 2025 Dataset

Building upon the foundational work established in VSASV 2023, the 2025 challenge introduced a significantly expanded and more sophisticated dataset design. While the 2023 challenge primarily focused on establishing baseline benchmarks with synthesized speech and replay attacks, the 2025 iteration adopted a more complex evaluation framework reflecting the evolving landscape of spoofing threats.

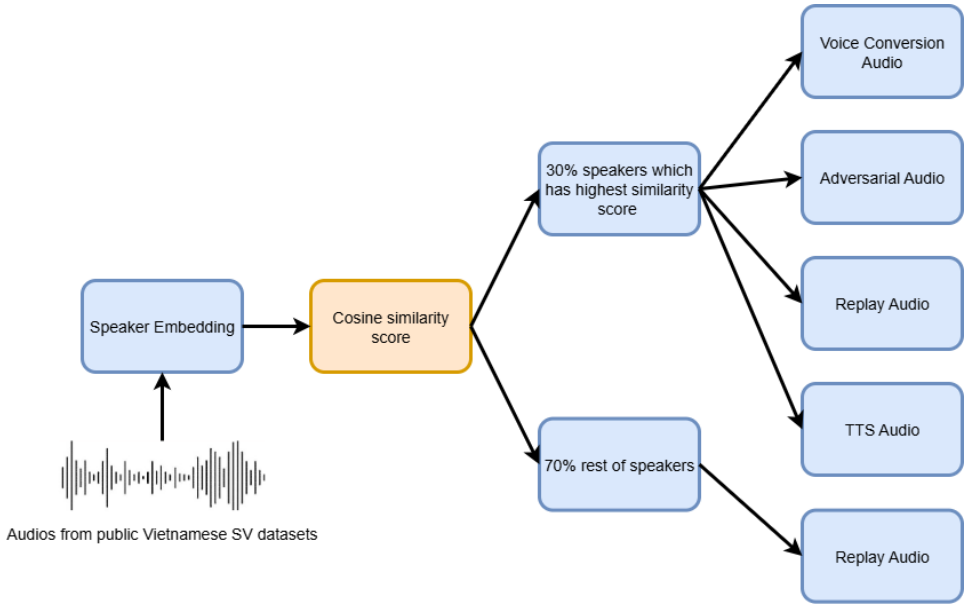

Figure 1. The pipeline of building VSASV challenge dataset [12].

3.1. Dataset Construction and Sources

The VSASV 2025 dataset was developed through the integration of multiple publicly available Vietnamese speech resources, specifically Vietnam-Celeb [9], VoxVietnam [13], and the VSASV corpus [14]. This multi-source approach ensured comprehensive coverage of speaker diversity, regional accent variations, and recording environment heterogeneity. Unlike the 2023 dataset which relied primarily on Vietnam-Celeb for bonafide samples, the 2025 challenge leveraged a broader collection of corpora to enhance acoustic variability and improve model generalization capabilities.

A fundamental principle maintained from the 2023 challenge was the strict speaker disjointness across training and evaluation partitions, preventing any speaker-level information leakage that could artificially inflate system performance. Table 5 presents the comprehensive statistics of the VSASV 2025 dataset across all partitions.

Compared to the 2023 challenge which provided approximately 226 hours of bonafide training data, the 2025 training set contained 71,617 bonafide utterances spanning 152.89 hours. While this represents a reduction in training data volume, it reflects a deliberate design choice to simulate realistic low-resource deployment scenarios more accurately. The evaluation sets, however, were substantially expanded: the public test set comprises 491,676 audio pairs (806.05 hours total), while the private test set contains 557,516 pairs (779.66 hours total), representing a significant increase over the 2023 evaluation protocol.

Table 5. Overall statistics of the VSASV 2025 dataset

Subset	Utterance Type	# Utterances	Duration (hours)
2*Train	Bonafide	71,617	152.89
	Spoofed	29,750	50.38
2*Public Test	Bonafide	235,577	353.29
	Spoofed	256,099	452.76
2*Private Test	Bonafide	185,577	276.42
	Spoofed	371,939	503.24

3.2. Spoofing Attack Distribution

A critical evolution from VSASV 2023 to 2025 lies in the strategic distribution of spoofing

attack types across dataset partitions. The 2023 challenge provided exposure to both synthesized speech (voice conversion) and replay attacks during training, with adversarial perturbations introduced at test time. In contrast, the 2025 challenge adopted a deliberately constrained training paradigm where only replay attacks were available during model development. This design decision was motivated by the realistic observation that in operational deployments, comprehensive labeled examples of all possible attack types are rarely available.

Table 6 details the distribution of spoofing techniques across dataset partitions. The training partition exclusively contains 29,750 replay attack utterances, establishing a known-attack baseline. The evaluation sets, however, introduce three categories of previously unseen attacks: voice conversion (VC), text-to-speech (TTS) synthesis, and adversarially perturbed samples.

Table 6. Distribution of spoofing attack types in VSASV 2025

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

The attack distribution reveals a strategic asymmetry between test sets. The public test emphasizes replay attacks (150,000 samples), maintaining consistency with the training distribution while introducing moderate quantities of unseen attack types. Conversely, the private test dramatically inverts this distribution, with voice conversion attacks dominating (210,139 samples) and replay attacks reduced to merely 5,000 samples. This inversion serves two evaluation objectives: first, assessing system robustness when the predominant test-time attack differs substantially from training conditions; second, evaluating generalization capability across spoofing methodologies that share no direct training examples.

The inclusion of adversarial attacks warrants particular attention. While the 2023 challenge applied adversarial perturbations to 50% and 100% of spoofed samples in public and private tests respectively, the 2025 challenge maintains substantial adversarial content (50,000 and 150,000 samples) but distributes it more strategically across different base attack types. This approach tests whether countermeasures can maintain robustness against perturbations applied to diverse spoofing foundations.

3.3. Speaker Similarity-Based Partitioning

A methodological innovation introduced in VSASV 2025 was the speaker similarity-driven data partitioning strategy, which represents a significant departure from the more conventional, randomly assigned speaker splits used in the 2023 edition. As illustrated in Figure 1, the process begins with the extraction of high-resolution speaker embeddings from all bonafide utterances using the ECAPA-TDNN architecture, a model known for its robustness in capturing fine-grained speaker-dependent acoustic cues. Pairwise cosine similarity scores are then computed across all available speaker combinations, producing a comprehensive similarity matrix that serves as the foundation for similarity-based stratification.

From this similarity distribution, speakers are partitioned into two disjoint subsets: the top 30% of most acoustically similar speaker pairs are assigned to the evaluation sets, while the remaining 70% form the training partition. This targeted allocation ensures that the verification trials in the evaluation phase are inherently challenging, as they predominantly involve speakers whose vocal timbres and prosodic characteristics closely resemble each other. Such high-similarity conditions are particularly demanding for both verification systems and spoofing countermeasures, pushing them beyond the relatively lenient conditions of prior challenges.

To further compound the difficulty, spoofed utterances corresponding to the evaluation speakers are generated using a wide spectrum of advanced and emerging attack methodologies. These include neural voice conversion systems capable of producing timbre-aligned conversions, state-of-the-art neural TTS synthesizers, and adversarial perturbation techniques that subtly manipulate audio waveforms to induce model failures. Traditional replay attacks are also incorporated, ensuring a comprehensive and realistic threat landscape that mirrors adversarial strategies observed in real-world deployments.

This similarity-stratified design directly addresses a key limitation in the VSASV 2023 challenge, where speaker distributions were largely random and thus failed to systematically emphasize high-confusability scenarios. By intentionally concentrating evaluation on acoustically similar speakers, the VSASV 2025 challenge introduces a more principled and stress-tested benchmarking framework. Moreover, this methodology is consistent with the low-resource training paradigm: systems must learn from a broad and acoustically diverse pool of training speakers but ultimately generalize to highly similar, previously unseen speakers who exhibit overlapping acoustic signatures. This creates an evaluation environment that not only reflects realistic adversarial conditions but also encourages the development of more robust, generalizable, and spoof-aware verification systems.

4. Evaluation Methodology

4.1. Performance Metrics

Both VSASV 2023 and 2025 challenges employ the Equal Error Rate (EER) as the primary evaluation metric for assessing system performance. The EER provides a scalar measure of system accuracy by identifying the operating threshold at which two types of errors occur with equal frequency: False Acceptance Rate (FAR) and False Rejection Rate (FRR). Systems achieving lower EER values demonstrate superior discrimination capability between legitimate target trials and various forms of non-target attempts.

The spoofing-aware speaker verification task is formulated as a binary classification problem where systems must distinguish between two trial categories. Target trials, labeled as positive (Label 1), consist of enrollment and test utterances that are both bonafide and originate from the same speaker. Non-target trials, labeled as negative (Label 0), encompass all remaining scenarios including: (1) bonafide impostor trials where both utterances are genuine but from different speakers, and (2) spoofed trials where the enrollment is bonafide but the test utterance has been manipulated through synthesis, conversion, replay, or adversarial perturbation.

The EER computation involves analyzing the distribution of system scores across target and non-target trial populations. For a given decision threshold θ, the system accepts trials with scores exceeding this threshold and rejects those below it. The False Acceptance Rate quantifies the proportion of non-target trials incorrectly accepted:

$$FAR(\theta) = \frac{\text{\# non-target trials with score} > \theta}{\text{Total \# non-target trials}} \tag{1}$$

Conversely, the False Rejection Rate measures the proportion of legitimate target trials incorrectly rejected:

$$FRR(\theta) = \frac{\text{\# target trials with score} \leq \theta}{\text{Total \# target trials}} \tag{2}$$

The EER is determined as the error rate at the threshold θ^* where these two quantities are equal:

$$EER = FAR(\theta^*) = FRR(\theta^*), \tag{3}$$

This metric provides a balanced assessment of system performance that does not favor either

conservative (high rejection rate) or permissive (high acceptance rate) operating points. In practical deployment scenarios, the actual operating threshold may differ from θ^* based on application-specific security requirements and usability constraints. However, EER serves as a standardized comparison metric across different systems and challenges.

The evaluation framework treats all non-target trials uniformly in the primary EER computation, reflecting the operational requirement that systems must reject any trial that is either from a different speaker or involves spoofed audio, regardless of the specific failure mode. This unified treatment emphasizes end-to-end system robustness under the combined threat model of conventional impostor attacks and various spoofing techniques. A system's ability to maintain low EER indicates successful integration of both speaker verification and spoofing countermeasure components, as failures in either dimension directly contribute to elevated error rates.

5. Baseline Systems

The VSASV challenges provided baseline systems to establish performance benchmarks and facilitate participant entry. While both challenges adopted a modular architecture separating speaker verification and spoofing detection components, the baseline configurations evolved significantly between iterations, particularly regarding the use of pre-trained models.

5.1. VSASV 2023 Baseline Configuration

The VSASV 2023 Challenge established its baseline system using two primary components trained from scratch on the provided data. For speaker verification, the baseline employed the ECAPA-TDNN [15] architecture, which had demonstrated strong performance in previous speaker recognition benchmarks. ECAPA-TDNN utilizes a Res2Net [16] backbone augmented with Squeeze-and-Excitation blocks [17] for channel-wise attention, combined with statistics pooling to aggregate frame-level features into utterance-level speaker embeddings. The model processes 80-dimensional Mel-filterbank features as input and applies data augmentation through room impulse responses [18] and additive noise from the MUSAN corpus [19] to improve robustness.

For spoofing detection, the 2023 baseline utilized the AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal graph attention networks) architecture, which had achieved competitive results in ASVspoof challenges. AASIST employs graph attention mechanisms to model relationships between spectral and temporal acoustic patterns, enabling effective discrimination between bonafide and spoofed speech across diverse attack types.

A critical constraint in VSASV 2023 was the prohibition of pre-trained models for the spoofing detection component. Participants were required to train countermeasure systems exclusively on the provided challenge data, preventing reliance on large-scale pre-training from external corpora. This restriction aimed to evaluate model architectures and training strategies under realistic low-resource conditions where extensive labeled spoofing data may not be available. The speaker verification component, however, could leverage standard data augmentation techniques and training recipes from open-source implementations.

The baseline system integrated these two components through score-level fusion, combining ASV similarity scores with CM detection scores to produce final verification decisions. This simple fusion approach provided a straightforward benchmark without requiring joint training or complex score calibration procedures.

5.2. VSASV 2025 Baseline Configuration

The VSASV 2025 Challenge introduced substantial modifications to the baseline architecture, most notably permitting the use of self-supervised speech foundation models for spoofing detection. This policy shift reflected the growing recognition that large-scale pre-trained representations, learned from diverse speech data, can provide robust acoustic features that generalize effectively to downstream anti-spoofing tasks, particularly under low-resource training conditions.

5.2.1. Speaker Verification Component

The ASV component maintained continuity with the 2023 baseline by employing the ECAPA-TDNN architecture. The model architecture remained largely unchanged, featuring Res2Net-based feature extraction with Squeeze-and-Excitation mechanisms [17] for adaptive channel recalibration. Multi-scale feature aggregation through channel- and context-dependent statistics pooling enables the model to capture both global speaker characteristics and local temporal patterns within utterances.

Input representations consist of 80-dimensional Mel-filterbank coefficients extracted from raw audio waveforms. Following feature extraction, frame-level embeddings are pooled and projected through fully connected layers to generate fixed-dimensional speaker embeddings. Cosine similarity between enrollment and test embeddings serves as the ASV score, with higher values indicating greater likelihood of same-speaker trials.

Training augmentation strategies mirror those from 2023, incorporating simulated room acoustics through impulse response convolution [18] and additive background noise from the MUSAN database [19]. These augmentations expose the model to acoustic variability during training, improving robustness to channel effects and environmental conditions encountered during evaluation. The implementation follows established open-source recipes to ensure reproducibility[2].

5.2.2. Spoofing Detection Component

The countermeasure baseline underwent significant architectural evolution in 2025, adopting the XLSR-Conformer + TCM framework introduced in recent anti-spoofing research [20]. This architecture leverages cross-lingual self-supervised representations (XLS-R) derived from wav2vec 2.0 [21] as a feature extraction backbone, marking a departure from the train-from-scratch constraint of 2023.

The XLS-R encoder, pre-trained on large-scale multilingual speech data exceeding 400,000 hours, generates contextualized representations that capture rich phonetic and acoustic information. For an input waveform O, the self-supervised module produces a sequence of high-dimensional feature vectors:

$$X = \mathrm{SSL}(O) = \{x_t \in \mathbb{R}^D \mid t = 1, \ldots, T\} \quad (4)$$

where D represents the embedding dimensionality and T denotes the temporal sequence length.

These pre-trained features undergo dimensionality reduction through a linear projection followed by scaled exponential linear unit (SeLU) activation:

$$\tilde{X} = \mathrm{SeLU}(\mathrm{Linear}(X)) \quad (5)$$

The transformed representations are subsequently processed through stacked Conformer blocks, which integrate multi-head self-attention (MHSA) mechanisms with convolutional modules. This hybrid architecture captures both long-range temporal dependencies through attention and local spectro-temporal patterns through convolution. A learnable class token, prepended to the feature sequence, accumulates global information across time

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

steps, with its final state serving as input to a binary classifier distinguishing bonafide from spoofed utterances.

The key innovation in this baseline lies in the Temporal-Channel Modeling (TCM) mechanism, which enhances the MHSA module by explicitly modeling interactions between temporal dynamics and channel-wise features. This cross-dimensional fusion enables the model to jointly reason about when specific acoustic patterns occur and which frequency channels exhibit anomalous characteristics indicative of synthesis artifacts. Empirical results on ASVspoof 2021 benchmarks demonstrated that XLSR-Conformer+TCM achieves a 26% relative error reduction compared to models without TCM [20], validating its effectiveness for detecting neural vocoder-based synthesis. Implementation details are available through open-source repositories[3].

5.2.3. System Integration

Following the SASV 2022 Challenge framework [8], the 2025 baseline adopts score-sum fusion to integrate ASV and CM components. This parameter-free approach combines normalized scores from both subsystems without requiring additional training or learned fusion weights. Specifically, ASV scores are computed as cosine similarities between speaker embeddings, while CM scores are passed through softmax normalization to map logits into the $(0, 1)$ probability range. The final decision score is obtained through simple arithmetic addition of these normalized values.

This straightforward fusion strategy provides interpretability and simplicity, enabling participants to focus on improving individual subsystem performance before exploring more sophisticated integration methods such as learned fusion networks or joint optimization frameworks.

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

5.3. Evolution and Rationale

The transition from VSASV 2023 to 2025 baseline systems reflects evolving perspectives on the role of pre-trained models in low-resource spoofing detection. The 2023 prohibition on pre-training emphasized architectural innovation and training efficiency under data scarcity, encouraging participants to develop models that could learn robust representations from limited labeled spoofing examples.

However, the 2025 policy shift recognizing self-supervised speech models acknowledges practical realities of modern anti-spoofing deployment. Foundation models pre-trained on massive unlabeled speech corpora have demonstrated remarkable transfer learning capabilities, often exceeding the performance of task-specific architectures trained from scratch on limited data. By permitting these models in 2025, the challenge aligned with contemporary best practices while maintaining the low-resource training constraint—the challenge data remained limited, but participants could leverage universal acoustic representations learned from broader speech distributions.

This evolution does not diminish the value of the 2023 approach but rather expands the solution space to include transfer learning strategies alongside architectural and algorithmic innovations. The comparative analysis of systems across both challenges provides valuable insights into the relative contributions of model architecture, pre-training strategies, and task-specific optimization in achieving robust spoofing-aware verification under resource constraints.

6. Challenge Results and Analysis

6.1. VSASV 2023 Challenge Results

The inaugural VSASV 2023 Challenge attracted significant participation from the research community, with 35 teams initially registering for the competition. Of these,

28 teams proceeded to sign data access agreements, indicating serious engagement with the challenge. However, actual submission rates revealed the inherent difficulty of the task: 12 teams submitted results for the public test evaluation, with only 7 teams proceeding to submit final systems for the private test set.

6.1.1. Public Test Performance

Table 7 presents the performance rankings on the VSASV 2023 public test set. The evaluation phase generated approximately 900 submissions across all participating teams, demonstrating active system development and iterative refinement during the competition period. Performance varied considerably across submissions, with EER values spanning from 2.60% for the top-performing system to 22.87% for the lowest-ranked submission.

Table 7. VSASV 2023 public test results ranked by EER performance

Rank	Team ID	EER (%)
1	TBQ	2.60
2	SpoofySV	2.86
3	Unknown	3.15
4	NNDam	3.77
5	VC-ML	3.86
6	AASR	4.08
7	HanoiVoice	5.64
8	Alpaca	13.43
9	Hynguyenthien	14.79
10	Kietha	15.46
11	Team008	20.14
12	FFYYTT	22.87

The public test results revealed a clear performance stratification, with the top six systems achieving EER below 5%, suggesting successful integration of speaker verification and spoofing detection capabilities. The substantial performance gap between rank 7 (5.64%) and rank 8 (13.43%) indicates a qualitative difference in system design or training methodology, separating highly competitive solutions from baseline-level implementations.

6.1.2. Private Test Performance and Final Ranking

Following public test evaluation, organizers conducted technical review of submitted systems, assessing both quantitative performance and solution methodology. The final ranking, based on private test results and system descriptions, revealed interesting dynamics in system generalization. Table 8 presents the official final standings.

Table 8. VSASV 2023 final ranking based on private test evaluation

Rank	Team ID	EER (%)
1	TBQ (NamiTech)	2.97
2	VC-ML (VCCorp)	9.89
3	Unknown (HUST)	21.17
4	Team008	21.46
5	Alpaca	25.17
6	SpoofySV	28.49
7	AASR	35.46

The TBQ system from NamiTech maintained its leading position across both evaluation phases, achieving 2.97% EER on the private test set. This consistency demonstrates robust generalization to the more challenging private evaluation conditions, where adversarial attacks were applied to 100% of spoofed utterances compared to only 50% in the public test. The slight performance degradation from 2.60% to 2.97% indicates effective countermeasures against adversarial perturbations.

Notably, ranking reshuffling occurred between public and private evaluations. The SpoofySV team, which achieved second place (2.86%) on public test, dropped to sixth position (28.49%) on private test, representing nearly a 10-fold performance degradation. This dramatic shift suggests potential overfitting to public test characteristics or vulnerability to the increased adversarial attack intensity in private evaluation. Conversely, Team008 demonstrated improved relative standing, advancing from 11th to 4th position despite absolute performance decline.

6.1.3. Top System Architectures

Table 9 summarizes the architectural choices and design strategies employed by the top three finalists. All leading systems adopted modular architectures separating speaker verification and spoofing detection components, consistent with the SASV paradigm.

Table 9. System architectures and strategies of top-3 teams in VSASV 2023

Component	TBQ (1st)	VC-ML (2nd)	Unknown (3rd)
ASV Models	ResNetSE34V2 ECAPA-TDNN RawNet3	SincNet ECAPA-TDNN	ECAPA-TDNN RawNet3
CM Models	ResNetSE34V2	AASIST	AASIST S2pecNet
Loss Functions	ArcFace	AAM-Softmax	AAM-Softmax
Backend Scoring	Cosine, AS-Norm	DNN	Cosine, Euclidean
Score Fusion	Weighted Avg	-	DNN
VAD	No	No	Yes
Data Augmentation	Yes	Yes	Yes

The winning TBQ system employed an ensemble approach, integrating three distinct speaker verification architectures (ResNetSE34V2, ECAPA-TDNN, RawNet3) with a ResNetSE34V2-based countermeasure module. Notably, this team utilized the same ResNetSE34V2 architecture for both ASV and CM tasks, suggesting that unified architectural frameworks can effectively handle both speaker discrimination and artifact detection when trained with appropriate objectives. The system employed ArcFace loss for discriminative embedding learning, combined with AS-Norm (Adaptive Score Normalization) and weighted averaging fusion to integrate multi-system outputs.

The second-place VC-ML system adopted a more streamlined configuration, combining SincNet and ECAPA-TDNN for speaker verification with AASIST for spoofing detection. AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal graph attention networks) emerged as a popular choice among multiple teams due to its demonstrated effectiveness in ASVspoof challenges. This system employed DNN-based backend scoring rather than simple cosine similarity, potentially enabling more sophisticated decision boundary learning.

The third-place Unknown team from HUST distinguished itself by incorporating Voice Activity Detection (VAD) as a preprocessing step, the only top-three system to do so. This design choice aims to remove non-speech segments that might introduce artifacts or confuse spoofing detectors. The team combined ECAPA-TDNN and RawNet3 for ASV with AASIST and S2pecNet for countermeasures, employing DNN-based fusion to integrate heterogeneous system outputs.

All top systems emphasized data augmentation as a critical training strategy, reflecting the consensus that acoustic variability and domain robustness are essential for effective spoofing-aware verification. The prohibition on pre-trained models in 2023 meant these systems relied entirely on architectural innovation, training optimization, and data augmentation to achieve competitive performance under low-resource conditions.

6.2. VSASV 2025 Challenge Results

The VSASV 2025 Challenge observed different participation dynamics compared to its predecessor. From 30 registered teams, 9 teams submitted valid results for the public test evaluation, with 8 teams proceeding to the final private test submission. While absolute participant numbers decreased from 2023, the submission quality and performance levels demonstrated substantial advancement, likely attributable to the maturation of spoofing-aware verification methodologies and the availability of pre-trained foundation models.

6.2.1. Public Test Performance

Table 10 presents the public test rankings for VSASV 2025. The evaluation revealed a considerably more challenging task compared to 2023, with EER values ranging from 20.79% to 36.85%. Even the top-performing system

achieved an EER nearly eight times higher than the 2023 public test leader, reflecting the increased difficulty introduced by the expanded attack diversity, speaker similarity-based partitioning, and low-resource training constraints.

Table 10. VSASV 2025 public test results ranked by EER performance

Rank	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

Despite the elevated absolute EER values, 9 out of 10 systems surpassed the official baseline performance (33.88%), demonstrating that participants successfully developed effective solutions beyond the provided reference implementation. The top-ranked SA-SVBK system achieved a relative improvement of 38.6% compared to the baseline, indicating substantial innovation in system design, training strategies, or data augmentation techniques.

Performance distribution in 2025 exhibited greater compression than 2023, with the gap between first and ninth place spanning only 13.09 percentage points (20.79% to 33.88%) compared to 20.27 percentage points in 2023 (2.60% to 22.87%). This compression suggests that the 2025 challenge established a more uniform difficulty level that challenged all participants relatively equally, preventing any single approach from achieving dominant superiority.

6.2.2. Private Test Performance

The private test evaluation introduced additional challenges through modified attack distributions, with voice conversion attacks dominating (210,139 samples) compared to the replay-heavy public test. Table 11 presents the final private test rankings.

Table 11. VSASV 2025 private test results and final ranking

Rank	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

Interestingly, ranking reversals occurred between public and private evaluations. The SV++ team [22], ranked fourth on public test (26.91%), achieved first place on private test (17.78%), demonstrating superior generalization to unseen attack distributions. This 9.13 percentage point improvement from public to private test represents remarkable robustness, particularly considering that the private test emphasized voice conversion attacks largely absent from training data.

Conversely, the SA-SVBK [23] team experienced a slight ranking decline from first to second place, though their absolute performance improved from 20.79% to 17.86%. This improvement alongside ranking loss indicates that SV++ achieved even greater gains under the modified evaluation conditions. The top two systems converged to nearly identical performance levels (17.78% vs. 17.86%), separated by only 0.08 percentage points, suggesting both teams developed highly robust solutions through different methodological approaches.

The leading system achieved a 51.5% relative improvement over the baseline (36.78%),

substantially exceeding the 38.6% relative gain observed on the public test. This acceleration of improvement from public to private evaluation suggests that top-performing systems incorporated design elements specifically targeting generalization across diverse attack types, rather than optimizing solely for the known attack distribution in public test.

Eight out of nine participating systems surpassed baseline performance on the private test, maintaining the high success rate observed in public evaluation. This consistency indicates that the challenge successfully motivated the development of genuinely robust spoofing-aware verification systems rather than solutions that exploit specific characteristics of the evaluation protocol.

6.2.3. Solution Analysis and System Architectures

Examination of submitted system descriptions reveals convergent design principles among top performers, as summarized in Table 12. All leading systems adopted modular fusion frameworks, separating speaker verification and countermeasure components before integrating their outputs at the score level.

Table 12. System architectures of participating teams in VSASV 2025

Rank	Team ID	ASV Module	CM Module
1	SV++	ERes2NetV2	XLSR-Conformer+TCM
2	SA-SVBK	MFA-Conformer	XLSR-Conformer+TCM
3	ChatJLPT	ECAPA-TDNN	AASIST
4	Brosh	ResNet-48	AASIST
5	RD	ECAPA-TDNN	AASIST
6	NLP Noobs	RawNet3	AASIST
7	Arrebol	ECAPA-TDNN	AASIST
8	Baseline	ECAPA-TDNN	XLSR-Conformer+TCM
9	TQ	ECAPA-TDNN	XLSR-Conformer+TCM

A clear architectural pattern emerged distinguishing top-two finishers from other participants: both employed self-supervised learning-based countermeasures (XLSR-Conformer+TCM), while most lower-ranked systems utilized AASIST architectures trained from scratch. This observation suggests that foundation models pre-trained on large-scale speech data provide significant advantages for spoofing detection under low-resource training conditions, particularly when generalizing to unseen attack types like the voice conversion attacks dominating the private test.

The top-ranked SV++ system combined ERes2NetV2 [24] for speaker verification with XLSR-Conformer+TCM for countermeasures. ERes2NetV2 represents an enhanced version of Res2Net architectures incorporating efficient residual connections and multi-scale feature aggregation, providing robust speaker embeddings. Notably, this team emphasized extensive data augmentation covering all three primary spoofing types (voice conversion, TTS, adversarial), despite these attack types being absent or underrepresented in the official training data. This augmentation strategy likely contributed to the system's exceptional private test performance, where unseen attack types predominated.

The second-place SA-SVBK system distinguished itself through the adoption of MFA-Conformer [25] (Multi-scale Feature Aggregation Conformer) for speaker verification, paired with XLSR-Conformer+TCM for spoofing detection. Rather than pursuing aggressive data augmentation, this team focused on data quality refinement, employing DBScan-based clustering to identify and remove noisy or potentially mislabeled training samples. This data cleaning approach aimed to improve training stability and representation quality by ensuring model optimization focused on clean, reliable examples.

The contrasting strategies of the top two systems—one emphasizing data diversity through augmentation, the other emphasizing data quality through cleaning—both proved highly effective, converging to nearly identical final performance. This convergence suggests that spoofing-aware verification under low-resource conditions admits

multiple successful solution pathways, and that data-centric approaches (whether increasing diversity or improving quality) may be as important as architectural selection.

Lower-ranked systems predominantly employed ECAPA-TDNN or similar architectures for speaker verification, paired with AASIST for countermeasures. Several teams (ChatJLPT, RD, Arrebol) adopted identical architectural combinations yet achieved divergent performance outcomes, with EER ranging from 24.37% to 32.48%. This performance variance among architecturally equivalent systems underscores the critical importance of training procedures, hyperparameter configurations, data preprocessing pipelines, and implementation details that extend beyond model selection.

The baseline system itself utilized ECAPA-TDNN with XLSR-Conformer+TCM, indicating that access to foundation models alone does not guarantee competitive performance. The substantial gap between baseline and top systems (36.78% vs. 17.78%) demonstrates that effective integration, training optimization, and data strategy remain essential even when leveraging pre-trained representations.

6.3. Cross-Challenge Comparison and Evolution

Comparing results across VSASV 2023 and 2025 reveals both continuity and evolution in spoofing-aware speaker verification research for Vietnamese. The most striking difference lies in absolute performance levels: the 2023 top system achieved 2.97% EER, while the 2025 leader achieved 17.78% EER. This six-fold increase in error rate does not indicate regression in methodology but rather reflects fundamental differences in challenge design and evaluation difficulty.

The 2025 challenge incorporated multiple factors that elevated task complexity. First, the training data volume decreased from 225.68 hours of bonafide speech in 2023 to 152.89 hours in 2025, reducing the available supervision for speaker modeling. Second, training data contained only replay attacks, while evaluation included substantial voice conversion, TTS, and adversarial attacks, creating severe train-test mismatch. Third, the speaker similarity-based partitioning strategy ensured that evaluation focused on acoustically confusable speakers, inherently increasing verification difficulty. Fourth, the dramatic inversion of attack distributions between public and private tests (replay-dominated public test vs. VC-dominated private test) tested generalization under distributional shift.

Despite these challenges, the relative improvement margins in 2025 (51.5% over baseline) exceeded those in 2023 (approximately 38.6% over implicit baseline performance), suggesting that participants in 2025 developed more sophisticated solutions to overcome the heightened difficulty. The convergence of top systems to similar performance levels in both challenges (2.97% vs. 9.89% in 2023; 17.78% vs. 17.86% in 2025) indicates that multiple teams independently discovered near-optimal approaches given the constraints and evaluation criteria.

Architecturally, the shift from prohibiting to permitting pre-trained models fundamentally altered the solution landscape. In 2023, top systems emphasized ensemble approaches and multi-model fusion (e.g., TBQ's combination of three ASV architectures), attempting to capture diverse acoustic patterns through architectural variety. In 2025, top systems converged on foundation model-based countermeasures (XLSR-Conformer+TCM), with differentiation occurring primarily through data strategies rather than model architecture diversity.

The consistent success of score-level fusion across both challenges indicates that simple, interpretable integration methods remain effective for combining heterogeneous subsystems. Neither challenge saw dominant adoption of end-to-end joint training approaches,

possibly due to the difficulty of balancing speaker verification and spoofing detection objectives within a unified optimization framework, or due to the practical advantages of developing and debugging subsystems independently before integration.

Both challenges observed substantial ranking fluctuations between public and private evaluations, highlighting the persistent challenge of generalization in spoofing-aware verification. Systems that perform well on known evaluation conditions may fail to maintain superiority when confronting modified attack distributions or adversarial intensifications. This observation emphasizes the importance of robust validation strategies and conservative system design that prioritizes generalization over fitting to specific evaluation characteristics.

7. Discussion and Future Work

The VSASV Challenge was conceived to catalyze research on spoofing-aware speaker verification (SASV) for low-resource languages, with Vietnamese serving as a representative and impactful case study. In contrast to prior challenges that predominantly focused on English, the VSASV efforts - across both the 2023 and 2025 editions - directly confront the persistent lack of publicly accessible spoofing resources for Vietnamese. This shortage has historically constrained the development of robust ASV and CM systems in underrepresented linguistic contexts.

By constructing a dedicated Vietnamese corpus containing both bonafide speech and multiple categories of spoofed audio - including replay, voice conversion, text-to-speech synthesis, and adversarially manipulated samples - the challenge introduces a realistic and comprehensive evaluation environment. This benchmark enables the community to systematically assess spoofing-aware ASV systems under conditions that mirror real-world threats. The collective results from participating teams highlight that numerous systems achieved substantial improvements over the baseline, demonstrating that reliable ASV and CM performance is attainable even in data-scarce scenarios.

The outcomes further highlight the strengths of ASV and CM components. Fusion-based or joint architectures consistently delivered more stable performance across both public and private evaluation tracks, reinforcing the importance of integrating these subsystems rather than treating them as isolated modules. Additionally, the strong performance of systems leveraging speech foundation models suggests that methodologies originally developed for English deepfake detection can be successfully transferred to Vietnamese when paired with appropriate fine-tuning and augmentation strategies.

Moving forward, continued advancement in Vietnamese SASV research will depend on expanding both the breadth and depth of spoofing data. Increasing dataset scale, incorporating more sophisticated attack types, and improving coverage across diverse speaking styles will be essential for building resilient verification models. A particularly promising direction is the development of multi-genre VSASV datasets that not only include conversational or read speech but also extend to domains such as Vietnamese deepfake singing, expressive speech, and cross-accent generation. Singing voice deepfakes [26], in particular, present unique acoustic and stylistic characteristics that may challenge existing CM and ASV pipelines, offering a rich avenue for future investigation.

In parallel, future research should explore unified, end-to-end frameworks that jointly optimize ASV and CM objectives. Such models could reduce reliance on late-fusion heuristics while promoting more cohesive decision-making within a single architecture. Ultimately, we anticipate that the VSASV Challenge - by continuously evolving its dataset

design, evaluation protocols, and multi-genre coverage - will stimulate sustained progress in spoofing detection and speaker verification for low-resource languages, contributing to more equitable and globally reliable speech authentication technologies.

8. Conclusion

This paper presented a consolidated overview of the VSASV 2023 and 2025 Challenges, highlighting their collective contributions toward advancing spoofing-aware speaker verification for Vietnamese. As a tonal and low-resource language, Vietnamese poses unique challenges for both ASV and CM modeling, including high intra-speaker variability, dialectal differences, and limited availability of annotated spoofing data. The VSASV series directly confronts these obstacles by constructing dedicated Vietnamese corpora that encompass a broad spectrum of spoofing techniques - from replay and neural voice conversion to modern TTS and adversarial attacks. The diverse system submissions across both years demonstrate that robust spoofing-aware verification is achievable even in data-constrained settings, particularly through subsystem fusion, data-centric augmentation strategies, and the adoption of cross-lingual SSL speech models.

The stepwise evolution between the two challenge editions reveals deeper insights into generalization under realistic threat conditions. The 2025 challenge, in particular, underscores the difficulty of detecting unseen deepfake types and the importance of designing systems capable of withstanding severe distributional shifts. These findings reaffirm the broader relevance of VSASV as a benchmark not only for Vietnamese but also for other underrepresented languages facing similar resource limitations.

Looking forward, future research should prioritize expanding the scale, genre diversity, and spoofing coverage of Vietnamese datasets - including emerging domains such as deepfake singing and expressive or cross-accent speech, which introduce complex tonal and stylistic variations. Further progress will depend on exploring unified, end-to-end SASV architectures that jointly optimize speaker verification and spoofing detection, minimizing reliance on handcrafted fusion schemes. By continuing to grow and refine the VSASV framework, the community can move toward more secure, equitable, and linguistically inclusive voice authentication technologies equipped to combat the rapidly evolving landscape of speech deepfakes.

References

[1] Z. Saquib, N. Salam, R. Nair, N. Pandey, A. Joshi, A Survey on Automatic Speaker Recognition Systems, Vol. 123, 2010, pp. 134–145.

[2] Z. K. Anjum, R. K. Swamy, Spoofing and Countermeasures for Speaker Verification: A Review, in: 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 467–471.

[3] H. Azzuni, A. E. Saddik, Voice Cloning: Comprehensive Survey (2025).

[4] X. Tan, T. Qin, F. Soong, T.-Y. Liu, A Survey on Neural Speech Synthesis (2021).

[5] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, Z.-H. Ling, ASVspoof 2019: A Large-scale Public Database of Synthesized, Converted and Replayed Speech, Computer Speech Language, Vol. 64, 2020, pp. 101114.

[6] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, H. Delgado, ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection, in: 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 47–54.

[7] X. Wang, H. Delgado, H. Tak, J. weon Jung, H. jin Shim, M. Todisco, I. Kukanov, X. Liu,

M. Sahidullah, T. H. Kinnunen, N. Evans, K. A. Lee, J. Yamagishi, ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale, in: The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024), 2024, pp. 1–8.

[8] Jee-weon Jung and Hemlata Tak and Hye-jin Shim and Hee-Soo Heo and Bong-Jin Lee and Soo-Whan Chung and Ha-Jin Yu and Nicholas Evans and Tomi Kinnunen, SASV 2022: The First Spoofing-Aware Speaker Verification Challenge, in: Interspeech 2022, 2022, pp. 2893–2897.

[9] V. T. Pham, X. T. H. Nguyen, V. Hoang, T. T. T. Nguyen, Vietnam-Celeb: A Large-scale Dataset for Vietnamese Speaker Recognition, in: Interspeech 2023, 2023, pp. 1918–1922.

[10] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common Voice: A Massively-Multilingual Speech Corpus (2020).

[11] H.-T. Luong, H.-Q. Vu, A Non-expert Kaldi Recipe for Vietnamese Speech Recognition System, in: Y. Murakami, D. Lin, N. Ide, J. Pustejovsky (Eds.), Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 51–55.

[12] P. T. Dat, H. L. Vu, N. T. T. Trang, The Vietnamese Spoofing-aware Speaker Verification Challenge 2025: Summary and Results, in: L. C. Mai, N. T. M. Huyen, N. T. T. Trang (Eds.), Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing, Association for Computational Linguistics, Hanoi, Vietnam, 2025, pp. 71–77.

[13] H. L. Vu, P. T. Dat, P. T. Nhi, N. S. Hao, N. T. Thu Trang, 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), 2025, pp. 1–5.

[14] V. Hoang, V. T. Pham, H. N. Xuan, P. Nhi, P. Dat, T. T. T. Nguyen, VSASV: a Vietnamese Dataset for Spoofing-Aware Speaker Verification, in: Interspeech 2024, 2024, pp. 4288–4292.

[15] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, in: Interspeech 2020, 2020, pp. 3830–3834.

[16] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2Net: A New Multi-Scale Backbone Architecture, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 2, 2021, pp. 652–662.

[17] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[18] S. Arellano, C. Yeh, G. Bhattacharya, D. Arteaga, Room Impulse Response Generation Conditioned on Acoustic Parameters (2025).

[19] D. Snyder, G. Chen, D. Povey, MUSAN: A Music, Speech, and Noise Corpus (2015).

[20] D.-T. Truong, R. Tao, T. Nguyen, H.-T. Luong, K. A. Lee, E. S. Chng, Temporal-Channel Modeling in Multi-head Self-Attention for Synthetic Speech Detection, in: Interspeech 2024, 2024, pp. 537–541.

[21] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 12449–12460.

[22] P. V. Hoang, H. B. Thu, H. V. Khanh, SV++'s Vietnamese Spoofing-Aware Speaker Verification Systems for VLSP 2025, in: L. C. Mai, N. T. M. Huyen, N. T. T. Trang (Eds.), Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing, Association for Computational Linguistics, Hanoi, Vietnam, 2025, pp. 82–88.

[23] N. T. Trung, T. D. An, C. H. Viet, SVBK System Description to the VLSP 2025 Challenge on Vietnamese Spoofing-Aware Speaker Verification, in: L. C. Mai, N. T. M. Huyen, N. T. T. Trang (Eds.), Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing, Association for Computational Linguistics, Hanoi, Vietnam, 2025, pp. 78–81.

[24] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, S. Zhang, J. Li, ERes2NetV2: Boosting Short-Duration Speaker Verification Performance with Computational Efficiency, in: Interspeech 2024, 2024, pp. 3245–3249.

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

[26] A. Guragain, T. Liu, Z. Pan, H. Sailor, Q. Wang, Speech Foundation Model Ensembles for the Controlled Singing Voice Deepfake Detection (CTRSVDD) Challenge 2024, 2024, pp. 774–781.