SVBK System Description to the VLSP 2025 Challenge on Vietnamese Spoofing-Aware Speaker Verification

Nguyen Tran Trung[1], *To Duy An*[1], *Chu Hoang Viet*[1]

[1]HUSTEP Lab, SoICT Department, HUST

Abstract

This technical report describes the SVBK team's approach to the Vietnamese Spoofing-Aware Speaker Verification Challenge at VLSP 2025. Our system consists of two independently trained components: an Automatic Speaker Verification module and a Countermeasure module, whose outputs are fused at the score level to produce the final SASV decision. The method emphasizes independent optimization of both modules to leverage their complementary capabilities. Our submission ranks second in the challenge, achieving an Equal Error Rate of 17.86% on the private test set, according to the announcement of the VLSP organizers.

Index Terms: Spoof-aware Speaker Verification, Score-fusion, Low-resources, Spoof Countermeasures

1. Introduction

Automatic Speaker Verification (ASV) has become an increasingly important biometric technology, enabling secure and convenient identity verification across a wide range of applications, from mobile banking to access control. However, despite its progress and widespread adoption, ASV systems remain vulnerable to various spoofing attacks. Techniques such as voice conversion (VC), text-to-speech (TTS) synthesis, and other advanced speech generation methods can be exploited to mimic legitimate users, thereby undermining the reliability of ASV. A fundamental challenge in secure speaker verification is to address two complementary tasks simultaneously: detecting whether an input utterance is spoofed and verifying whether it originates from the claimed speaker. Traditional ASV systems are not inherently designed to handle spoofing attacks, while standalone countermeasure systems cannot determine speaker identity. This gap motivates the development of spoof-aware speaker verification (SASV), which seeks to integrate both capabilities into a unified framework.

Current solutions to SASV typically combine independent ASV and spoofing countermeasure (CM) systems. Fusion is performed at the score level [1, 2], the embedding level [3, 4], or through hybrid approaches that exploit both [5]. Each strategy offers distinct advantages and limitations: score-level fusion is computationally efficient and requires fewer resources, while embedding-based or hybrid fusion often yields superior discrimination by leveraging richer speaker and spoof representations.

In the deep learning context, however, the scale and quality of training data often exert a stronger influence on performance than architectural refinements or scoring techniques. Most existing studies are benchmarked on the SASV Challenge 2022 datasets [6], which integrate VoxCeleb2 [7] for ASV training and ASVspoof 2019 [8] LA for CM training. Although these large-scale corpora enable robust evaluation, only a limited number of works examine the generalization of proposed methods on smaller or less diverse datasets, where performance may degrade due to restricted speaker coverage or spoof variability.

In this study, we report our methods and results on the Vietnamese Spoof-Aware Speaker Verification (VSASV) track of the VLSP 2025 Challenge. Our best-performing system adopts a score-fusion framework, where the final decision is obtained via a weighted summation of two complementary scores: (i) the cosine similarity between the test and enrollment speaker embeddings, and (ii) the spoof detection score of the test utterance. On the other hand, to mitigate the effect of label noise in the organizer 's training dataset, we employ the DBSCAN clustering algorithm to identify and remove mislabeled samples, thereby refining the training set. With this approach, our system achieves an Equal Error Rate (EER) of 17.86% on the Private Test set, ranking second in the competition, only 0.08% higher than the top-performing system.

2. Methodology

2.1. Dataset & Cleaning

The dataset provided by the VLSP organizers comprises a total of 101,367 audio samples collected from 815 speakers, including 71,617 bonafide and 29,750 spoofed utterances. The dataset is divided into a training set and a testing set in a 9:1 ratio for experimental purposes. This ensures that the training data remains sufficiently sizable while preserving a significant fraction for performance assessment. This split is designed to maintain speaker diversity across both partitions as well as preserve the distribution balance between bona fide and spoofed samples.

In several instances, bona fide audio recordings with the same speaker label are actually uttered by different speakers, resulting in significant label noise in the training data, while ASV subsystem relies heavily on having cleaned labels. To mitigate this problem, we propose a DBSCAN-based dataset cleaning procedure applied individually to the embeddings of each speaker ID. In particular, an ASV model is first trained on the initial dataset, which extracts speaker embeddings for every bona fide utterance. Then, using DBSCAN with a threshold of 0.3, a minimum sample size of 3, and cosine distance as the similarity metric, these embeddings are clustered. Since the model is trained with the three sub-center ArcFace loss, we expect the embeddings of each speaker to naturally form up to three major clusters. This accounts for possible intra-speaker variations such as different recording environments (e.g., noise conditions, microphone types). Therefore, we keep utterances from each speaker's top three largest clusters to ensure that we preserve meaningful diversity while filtering outliers.

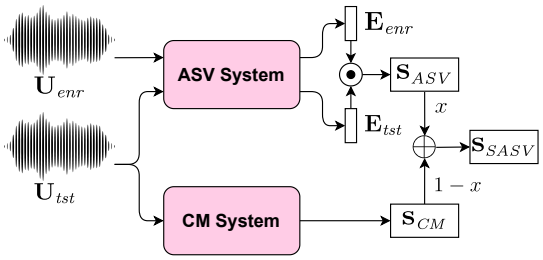

Figure 1: *Demonstration of the weighted-sum score fusion technique, where $\odot$ means dot-product, $\oplus$ means weighted addition.*

2.2. ASV subsystems

The goal of the ASV system is to determine whether a test utterance belongs to the claimed speaker. Modern ASV architectures typically consist of a frame-level feature extractor, a pooling layer, and an utterance-level feature extractor [9]. In our work, two model architectures are examined. The first, **MFA-Conformer** [10], introduces Multiscale Feature Aggregation (MFA) on top of the original Conformer encoder [11]. The second, **ResNet-SE**, constructs the encoder block-the central component of the network-by combining the Squeeze-and-Excitation module [12] with Residual Mappings [13].

Each ASV system produces a 192-dimensional speaker embedding vector, representing extracted features, for an input utterance. More specifically, for an input audio $\mathbf{U}$, a speaker embedding $\mathbf{E}$ is extracted as follows:

$$\mathbf{E} = f_{\text{ASV}}(\mathbf{U}) \tag{1}$$

where $f_{ASV}(\cdot)$ represents the ASV system.

For the ASV task, embeddings from the enrollment and test utterances are extracted, and their similarity is computed via the inner product, yielding a speaker similarity score that determines whether the two utterances originate from the same speaker.

2.3. CM subsystems

Spoof detection is a binary classification task in which the system determines whether a given utterance is bona fide or spoofed. For each test input, the model evaluates two possible hypotheses: the utterance is either genuine or the result of a spoofing attack.

For this challenge, we implement two model architectures for the countermeasure subsystem. The first is the XLSR-Conformer-TCM [14], a model that performs well on the ASVSpoof2021 benchmark [15]. The second model is an adaptation of the ResNet-SE encoder originally used in the ASV task. To leverage it for spoof detection, we append a fully connected layer on top of the embedding extractor. This additional layer outputs a one-dimensional vector containing two logits, each representing the probability that the input is spoofed or bona fide, where a higher logit value indicates greater confidence in the associated class. The following formula illustrates the spoof score extraction of a CM system:

$$\mathbf{o}_{\text{CM}} = f_{CM}(\mathbf{U}) \tag{2}$$

where $\mathbf{U}$ is the input utterance, $f_{\text{CM}}(\cdot)$ denotes the CM system, and $\mathbf{o}_{\text{CM}}$ is the final vector of two logits corresponding to the spoof and bona fide scores. By apply Softmax on the logits, the final CM score can be obtained by deriving the probability that an audio is bona fide.

For the ResNet-SE-based CM system, we utilize Log-Magnitude Spectrum features. This representation has proven to be effective in recent studies on spoofed audio detection [16].

2.4. Fusion techniques

Our primary results are obtained using the score-level weighted summation strategy. As illustrated in Figure 1, during inference two inputs are provided: the enrollment utterance $\mathbf{U}_{enr}$ and the test utterance $\mathbf{U}_{tst}$. Both are passed through the ASV system, which produces two embeddings, $\mathbf{E}_{enr}$ and $\mathbf{E}_{tst}$. The cosine similarity score $\mathbf{s}_{ASV}$ between these embeddings is then calculated via the dot product. At the same time, the CM system also processes $\mathbf{U}_{tst}$, extracting the corresponding spoofing score $\mathbf{s}_{CM}$. Finally, the two subsystem scores are combined using weighted summation to form the overall SASV decision, following the formula:

$$\mathbf{s}_{\text{SASV}} = x \cdot \mathbf{s}_{\text{CM}} + (1 - x) \cdot \mathbf{s}_{\text{ASV}} \tag{3}$$

where x is empirically determined to maximize performance on the public test set before being applied to the private one. Among all the methods, this proves to be the most effective.

Besides, we adopt a cascade strategy for score fusion. Since the two tasks are distinct and largely independent, it is reasonable to evaluate the audio separately on both tasks. In this method, the CM system evaluates the test audio and generates a score that is compared to a predetermined threshold σ. This process is illustrated as follows:

$$\mathbf{s}_{\text{SASV}} = \begin{cases} -1 & \text{if } \mathbf{s}_{\text{CM}} \geq \sigma \\ \mathbf{s}_{\text{ASV}} & \text{if } \mathbf{s}_{\text{CM}} < \sigma \end{cases} \tag{4}$$

Should the score fall below the bona fide cutoff, the system returns -1, rejecting the test audio as spoofed. This threshold is determined by optimization: we start with an initial threshold of 0.5 and then apply a local search to minimize the EER on the CM task. Otherwise, the system calculates the cosine similarity between the embeddings of the enrollment and test audio generated by the ASV subsystem.

Another idea of score-level fusion is to multiply the scores after raising $\mathbf{s}_{\text{CM}}$ to the power of q, which is referred to as Power Weighted Score Fusion (PWSF) [17]. The fusion function is expressed as:

$$\mathbf{s}_{\text{SASV}} = \mathbf{s}_{\text{ASV}} \cdot (\mathbf{s}_{\text{CM}})^q \tag{5}$$

where $\mathbf{s}_{\text{CM}}$ is normalized to the range $(0, 1)$ using a softmax function applied to the CM system output, and the probability of the bona fide class is used in the calculation."

Assuming that $q \geq 1$, this operation is valid when $\mathbf{s}_{\text{CM}}$ lies in the range $(0, 1)$ after the softmax operation. Multiplying the scores from the two subsystems is reasonable since only high values of both scores determine the identity of the test utterance in the context of the SASV task, which can also be interpreted as a logical AND operation. Raising $\mathbf{s}_{\text{CM}}$ to the power of q makes it more categorical while still continuous on the interval $(0, 1)$, in contrast to making a hard decision based on a fixed threshold.

3. Experiment Setup

3.1. ASV subsystems

We set up the training framework for the ASV subsystems in the same way as the repository that produced the results of the

Table 1: *EER (%) results on the Public Test and Private Test partition of VLSP 2025 evaluation.*

System	Fusion Technique	SPF-EER	SV-EER	SASV-EER	
		Public	*Public*	*Public*	*Private*
ResNet-SE (CM)	-	3.63%	-	-	-
XLSR-Conformer-TCM	-	**1.36**%	-	-	-
ResNet-SE (ASV)	-	-	**31.62**%	-	-
MFA-Conformer	-	-	31.72%	-	-
Baseline	-	-	-	24.21%	-
ResNet-SE (ASV) + Resnet-SE (CM)	Weighted Sum (x = 0.99)	-	-	20.79%	-
ResNet-SE (ASV) + XLSR-Conformer-TCM	Weighted Sum (x = 0.7)	-	-	19.71%	18.07%
MFA-Conformer + XLSR-Conformer-TCM	Weighted Sum (x = 0.75)	-	-	**19.6**%	**17.86**%

well-known ECAPA-TDNN[1] . The ASV subsystems are trained using our automatically cleaned dataset, which contains approximately 67k bona fide utterances from 741 speakers. Each audio file is randomly cropped into two-second segments to standardize input length. To enhance model robustness, we perform data augmentation by adding real-world noise from the MUSAN (music and noise) dataset and applying reverberation effects derived from the RIR dataset. The audio is subsequently transformed into 80-dimensional filterbank features, which serve as input to the ASV systems. Lastly, we apply SpecAugment [18] on the filterbanks, where random masking is applied to between 0 and 5 frames in the time dimension and 0 to 10 channels in the frequency domain.

The ASV subsystems are evaluated on the Public Test without spoofed audios.

3.2. CM subsystems

For the CM subsystems, the training framework is derived from the authors of the XLSR-Conformer-TCM[2] . We train the system using the entire training set provided by the VLSP organizers. The cleaned version of the dataset is not used, as its purpose was to remove incorrectly labeled bona fide samples caused by human error. Since the spoof data is generated according to the contest rules, it can be labeled correctly automatically, which we verified by manually inspecting several audio samples. During training, we extract random four-second segments from the original recordings, and all audio samples are augmented using the RawBoost technique [19]. The choice of four seconds is based on prior studies that report this duration to be effective for training CM models.

For the ResNet-SE-based CM model, the Log Magnitude Spectrum is employed as the primary input feature. The Short-Time Fourier Transform (STFT) parameters are configured as follows: a Fourier transform size of 512, a hop length of 160, and a window size of 400. Additionally, 60 mel-filter banks are applied for optional dimensionality reduction. The model is trained on an NVIDIA P100 GPU with a batch size of 32 for 10 epochs. The Adam optimizer is used with an initial learning rate of 0.001, and a scheduler decreases the learning rate by 20% every three epochs.

The CM subsystems are evaluated on the full Public Test, with target and non-target pairs labeled as bona fide.

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

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

3.3. SASV fusion

After obtaining individual scores for each utterance pair from the ASV and CM subsystems, we optimize the score-summation weights to minimize the Equal Error Rate (EER). Specifically, a linear search with a step size of 0.05 is conducted over the development set to determine the optimal weight configuration.

4. Results

4.1. Experiment results

Table 1 summarizes our results from the competition. As shown in the first two rows, the Conformer-based model demonstrates superior capability to detect spoofed audio, achieving a notably low equal error rate (EER) of 1.36%, compared to the ResNet-SE baseline. In contrast, the ASV systems exhibit considerably higher error rates on the test set: the ResNet-based system reaches an EER of 31.62%, which is only marginally better (by 0.1%) than the MFA-Conformer system. These unusually high error rates are attributed to label inconsistencies present in the competition's test set, where a substantial portion of the trials appear to be mislabeled. Together with the label noise previously identified in the training set, these inconsistencies substantially impair system performance—particularly given the relatively limited size of the dataset for training robust speaker verification models, even after applying our proposed label noise filtering strategy.

The last three rows of Table 1 present our score-fusion results and final submissions for the entire SASV task, evaluated on both the Public Test and Private Test sets of the competition. Using empirically optimized fusion weights, the weighted-sum results on the Public Test set show relatively modest differences across configurations. The best performance is achieved by combining two Conformer-based systems, resulting in an EER of 19.6%, which represents only a relative enhancement of 1.2% compared to the highest EER obtained from summing the ResNet-SE systems. However, this choice of system fusion also records an improvement of about 23.5% over the Baseline, which employs ECAPA-TDNN and XLSR-Conformer as two subsystems and averages the scores obtained from those systems. Under the same fusion strategy, the best-performing system on the Private Test set is the combination of our MFA-

Conformer and XLSR-Conformer-TCM models, which attains an EER of 17.86%.

Since the organizer gives us six chances for Private Test submission, most submissions are used to cross-check weighted sum on Public and Private Test, leading to most Private Test EER values being missing. While most of the promising results are achieved using the linear weighting technique, both the cascaded strategy and the PWSF method yield relatively poor performance on our evaluation set, with results that are unstable and consistently inferior to those of the linear fusion approach.

4.2. Observations

While training the ResNet-SE system, we notice that several spoofed samples exhibit distinctive spectral patterns that are visible even to the human eye. This observation motivates the use of logarithmic spectral representations for spoof detection using a lightweight model such as ResNet-SE. However, this design choice also explains why the system struggles to generalize effectively to unseen spoofing attacks in the public set.

The optimized fusion weights are also noteworthy, as they exhibit negligible variation and inconsistency across different model combinations. During our experiments, we perform a linear search within the range $(0, 1)$ to identify the optimal weights and observe a steady improvement in performance until convergence is reached. This behavior can be explained by the distinct characteristics of each subsystem: within a given task, different models produce scores that follow diverse distributions. When combined, these complementary score distributions interact in unique ways, leading to varying optimal weights across different fusion settings.

Although several promising results in the SASV Challenge 2022 were obtained through embedding-level fusion for final score computation, such approaches typically require post fine-tuning after pretraining each subsystem, making the overall process computationally expensive and time-consuming. Therefore, in this work, we restrict our experiments to score-level fusion, which prove to be more efficient while still yielding encouraging results in the competition.

5. Conclusion

In this paper, we present our solution to the VLSP 2025 Challenge on Vietnamese Spoofing-Aware Speaker Verification. Our solution comprises dataset cleaning using DBSCAN, an ASV system to handle the speaker verification task with the MFA-Conformer model, a CM system to handle the spoof detection task with the XLSR-Conformer model, and a comparison of various fusion techniques to return the best EER result. Using these techniques, the best result from our developed system ranks second place in the SASV task of the VLSP 2025 Challenge.

6. References

[1] Peng Zhang and Peng Hu and Xueliang Zhang, "Norm-constrained Score-level Ensemble for Spoofing Aware Speaker Verification," in *Interspeech 2022*, 2022, pp. 4371–4375.

[2] Y. Zhang, G. Zhu, and Z. Duan, "A probabilistic fusion framework for spoofing aware speaker verification," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 77–84.

[3] Jeong-Hwan Choi and Joon-Young Yang and Ye-Rin Jeoung and Joon-Hyuk Chang, "HYU Submission for the SASV Challenge 2022: Reforming Speaker Embeddings with Spoofing-Aware Conditioning," in *Interspeech 2022*, 2022, pp. 2873–2877.

[4] W. Ge, H. Tak, M. Todisco, and N. Evans, "Can spoofing countermeasure and speaker verification systems be jointly optimised?" 2023. [Online]. Available: https://arxiv.org/abs/2303.07073

[5] Haibin Wu and Lingwei Meng and Jiawen Kang and Jinchao Li and Xu Li and Xixin Wu and Hung-yi Lee and Helen Meng, "Spoofing-Aware Speaker Verification by Multi-Level Fusion," in *Interspeech 2022*, 2022, pp. 4357–4361.

[6] J. weon Jung, H. Tak, H. jin Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-J. Yu, N. Evans, and T. Kinnunen, "Sasv 2022: The first spoofing-aware speaker verification challenge," 2022. [Online]. Available: https://arxiv.org/abs/2203.14732

[7] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018*. ISCA, Sep. 2018. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1929

[8] T. Massimiliano, W. Xin, V. Ville, M. Sahidullah, D. Héctor, N. Andreas, Y. Junichi, E. Nicholas, T. H. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech 2019*, 2019, pp. 1008–1012.

[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[10] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-y. Lee, and H. Meng, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *Proc. Interspeech*, 2022, pp. 306–310.

[11] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 630–645.

[14] D.-T. Truong, R. Tao, T. Nguyen, H.-T. Luong, K. A. Lee, and E. S. Chng, "Temporal-channel modeling in multi-head self-attention for synthetic speech detection," in *Proc. Interspeech*, 2024, pp. 537–541.

[15] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "Asvspoof 2021: Accelerating progress in spoofed and deepfake speech detection," pp. 47–54, 2021.

[16] A. Alenin, N. Torgashov, A. Okhotnikov, R. Makarov, and I. Yakovlev, "A subnetwork approach for spoofing aware speaker verification," in *Proc. Interspeech*, 2022, pp. 2888–2892.

[17] A. Aliyev and A. Kondratev, "Intema system description for the asvspoof5 challenge: Power weighted score fusion," in *Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, 2024, pp. 152–157.

[18] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[19] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6382–6386.