SV++'s Vietnamese Spoofing-Aware Speaker Verification Systems for VLSP 2025

Pham Viet Hoang **Ho Bao Thu** **Ha Viet Khanh**
HuSTeP Lab, Hanoi University of Science and Technology
{hoang.pv224854, thu.hb226003, khanh.hv225979}@sis.hust.edu.vn

Abstract

This paper presents our system for the Vietnamese Spoofing-Aware Speaker Verification in VLSP 2025 challenge. The proposed system consists of an automatic speaker verification sub-system, a spoof detection sub-system, and a fusion module operating at either the score or embedding level. To overcome limited model generalization caused by insufficient data, we employed augmentation strategies such as adversarial perturbation, text-to-speech synthesis, and voice conversion. Our system achieved the best performance with EERs of 19.84% and 17.78% on the public and private test datasets respectively, ranking first in the VLSP 2025 challenge.

Keywords: Vietnamese Spoof-aware Speaker Verification, spoof detection, data augmentation.

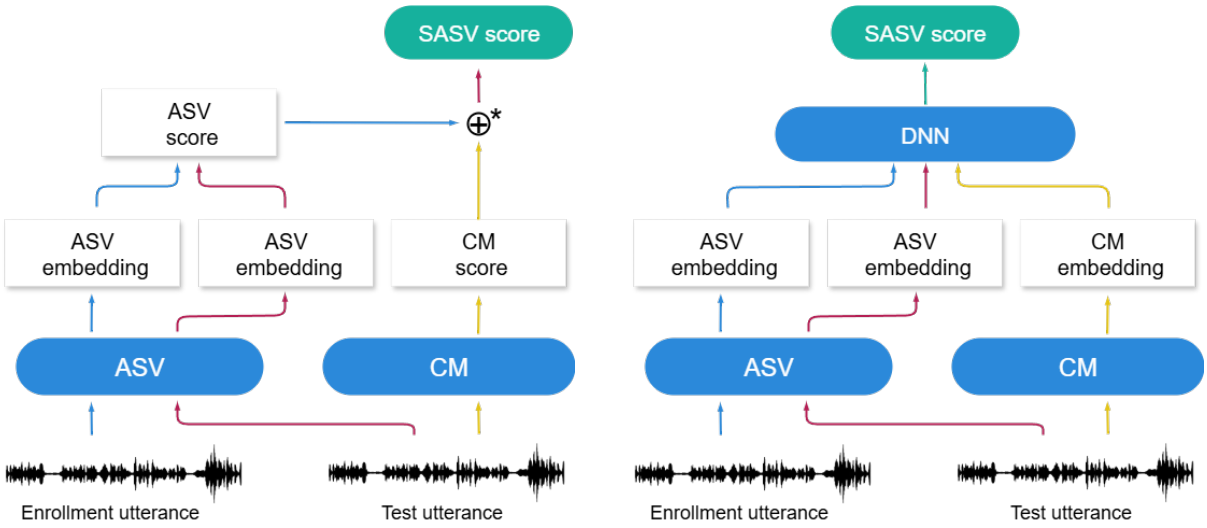
1 Introduction

Automatic speaker verification (ASV) has become increasingly important in security-critical applications such as voice biometrics and fraud prevention. However, the rapid development of voice synthesis and conversion technologies has introduced severe vulnerabilities. Malicious actors can now generate highly realistic audio deepfakes, threatening the reliability of ASV systems and enabling misinformation or identity fraud.

Recent surveys [1, 2] have reviewed the evolution of speech deepfake detection from handcrafted features and traditional models to deep learning, self-supervised methods, and end-to-end systems. An effective anti-spoofing system often relies on the integration of ASV and countermeasure (CM) sub-systems. The ASV sub-system is responsible for verifying a speaker's claimed identity based on their voice, with state-of-the-art models such as ECAPA-TDNN [3], MFA-TDNN [4], and Res2Net-based variants [5, 6]. The CM sub-system, on the other hand, detects and rejects spoofed speech, where leading approaches are AASIST [7], RawNet-based models [8], and recent transformer-based architectures [9]. Integration pipelines can be categorized into four types. Cascaded systems begin with an ASV classifier followed by the CM detector. Score-level fusion combines the output scores from both ASV and CM models to form a unified decision while embedding-level fusion integrates ASV and CM representations, either by concatenating their embeddings or extracting a joint representation. Finally, integrated end-to-end systems learn a unified Spoofing-Aware Speaker Verification (SASV) embedding directly, without separate ASV and CM modules. These surveys further indicate that score-level and embedding-level fusion consistently achieve superior performance, motivating this work to focus on enhancing these two strategies.

The Vietnamese Language and Speech Processing (VLSP) 2025 evaluation campaign features several shared tasks on text and speech processing. Among them, the Vietnamese Spoofing-Aware Speaker Verification Challenge (VSASV) aims to advance research in ASV and spoof detection for Vietnamese, where resources remain limited. The challenge provides a Vietnamese speech dataset collected from multiple speakers and emphasizes generalization by evaluating models on unseen speakers. The training partition comprises 815 speakers and 101,367 utterances, totaling 203.28 hours. Among these, 71,617 are bonafide samples, while the remainder are labeled as spoofed. This imbalance may hinder generalization across diverse spoofing scenarios, underscoring the need for augmentation. Among various attack sources referred to in [1], adversarial perturbations, speech synthesis, and voice conversion are the most prominent; we therefore adopt them as the main augmentation approaches.

This paper is organized as follows: Section 2 outlines the methodology; Section 3 describes the



(a) Weighted-sum fusion

(b) fusion based on Deep Neural Networks (DNN)

Figure 1: Overall architectures of the two fusion strategies: (a) weighted-sum fusion, where $\oplus^*$ denotes a linear interpolation with coefficient α; (b) DNN-based fusion.

data augmentation, implementation, and results; and Section 4 concludes the work.

2 Methodologies

2.1 Overall architecture

The overall architecture of our system extends two baseline solutions introduced in the SASV 2022 challenges [10]: (1) a score-level fusion combining the ASV cosine similarity and CM output scores, and (2) an embedding-level fusion using a deep neural network (DNN) backend classifier. In both approaches, the ASV and CM sub-systems are first trained independently and subsequently integrated into the overall system.

2.2 ASV sub-system

We evaluate three representative architectures: ECAPA-TDNN [3], MFA-TDNN [4], and ERes2NetV2 [6]. ECAPA-TDNN enhances the traditional TDNN [11] with channel attention and Res2Net modules to better capture speaker characteristics. MFA-TDNN further enhances ECAPA-TDNN by incorporating multi-scale feature aggregation to enrich the speaker embeddings. Meanwhile, ERes2NetV2 builds on Res2Net [5] with strengthened residual connections, enabling more efficient representation learning.

2.3 CM sub-system

The CM sub-system adopts the XLSR-Conformer architecture with Temporal-Channel Modeling (TCM) [9]. While XLSR-Conformer leverages a pre-trained multilingual SSL model with a Conformer backbone to capture both local and global dependencies in speech signals, TCM further exploits correlations between temporal and spectral domains, motivated by the conjecture that synthetic speech artifacts often occur in specific time-frequency regions [12, 13]. Further implementation details for optimization are provided in Section 3.2.

2.4 SASV system

Once the ASV and CM sub-systems are independently trained, they are frozen and used to extract scores or embeddings. These outputs are then integrated at the score level or the embedding level.

At the score level, fusion is performed via linear interpolation. The ASV branch computes a cosine similarity score between enrollment and test embeddings, while the CM branch outputs a spoofing detection score for the test utterance. These two scores are then linearly combined with tunable weights, where the weighting factor reflects the relative importance of each sub-system. The overall architecture of Baseline 1 is illustrated in Figure 1a.

At the embedding level, we employ a learnable back-end classifier. It operates on three embeddings: a pair of ASV embeddings from the enrollment and test utterances, and a CM embedding from the test utterance. These representations are concatenated and then passed through a feed-forward DNN with multiple fully connected layers using LeakyReLU [14] activations, followed by a

final linear layer that predicts bonafide vs. spoofed trials. This design enables the network to jointly model speaker similarity and spoofing cues within a unified representation space. The overall architecture of Baseline 2 is illustrated in Figure 1b.

3 Experiments and Results

3.1 Data augmentation

One of the major challenges we identified in the VLSP 2025 VSASV dataset is the imbalance of spoofed audio in the training data compared to the public test set. Notably, spoofed utterances comprise less than 30% of the training data, while they constitute approximately 52% of the public test set. Since the spoofing techniques in the test remain undisclosed, we did not attempt to replicate them directly. Instead, we explored diverse augmentation strategies to enrich synthetic spoof samples. An overview of the original VLSP 2025 VSASV dataset together with our augmented data is presented in Table 1.

Table 1: Statistics of the training data (* describes the spoofed data generated by our team; the remaining data were provided by the competition organizers)

Type	#Utterances	#Hours
Bonafide	71,617	152.89
Spoof	29,750	50.39
Voice conversion*	6,180	13.53
Adversarial*	11,071	24.00
Text-to-speech*	2,069	2.84
All	**120,687**	**243.65**

Text-to-Speech (TTS)

For TTS augmentation, we utilized the pretrained `F5-TTS-Vietnamese-100h` model [15], released on Hugging Face. Transcripts were obtained by running a pretrained Chunkformer[1] [16] on the training audio, from which we retained only utterances longer than two seconds to ensure sufficient duration for reliable synthesis. The system occasionally produced code-switched segments that degraded TTS quality, these were manually filtered out. To increase variability, the dataset was generated with two different speaking rate configurations-normal speed (1.0x) and slower speed (0.7x)- using non-overlapping training data partitions. This procedure yielded 2,069 utterances totaling 2.84 hours, capturing human-like articulation and prosodic variation to enhance spoof detection robustness.

Adversarial perturbation (AP)

Following [17], we adopted the Double Deceiver framework [18] to generate adversarial spoofing data. An ECAPA-TDNN and an XLSR Conformer combined with a TCM module were first trained on the original training set as base models. These models were then used as targets for generating adversarial attacks. Prior to attack generation, the reference recordings were normalized to -3dB to standardize signal levels, followed by enhancement with DeepFilterNet 3 [19] to reduce noise and improve clarity. The Double Deceiver then optimized perturbations against both target networks simultaneously, producing adversarial examples with a higher likelihood of generalizing beyond a single model. In total, this process yielded 11,071 utterances with a duration of 24.00 hours.

Voice conversion (VC)

To further expand the spoofing diversity, we incorporated VC augmentation. In this step, we employed the Retrieval-based Voice Conversion (RVC) Project[2], a publicly available toolkit that enables fine-tuning on new speakers with relatively small amounts of data. RVC integrates a pretrained UVR5 model for fast vocal-instrument separation and a HuBERT-based [20] representation model for effective voice conversion. Given that the training set contained many noisy recordings, we manually selected a subset of speakers with the cleanest speech and applied the same enhancement as in the adversarial preprocessing before RVC training. For each selected speaker, we trained a dedicated VC model and then applied it to convert audio from all other speakers, thereby creating cross-speaker spoof samples. With clean speaker data and the low data requirements of RVC, this method proved to be particularly efficient. This process resulted in 6180 high-quality VC spoofed utterances derived from over 1000 bonafide audio files across more than 20 speakers. By introducing speaker-mimicked spoofing conditions, VC augmentation provided an additional layer of variety that complements both TTS-generated and adversarially crafted spoof data.

Through this augmentation pipeline, we significantly increased both the quantity and variety of spoofed training data. The inclusion of synthetic

[1] https://github.com/khanld/chunkformer

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

speech from TTS, speaker-altered audio from VC, and adversarially crafted attacks enabled us to better approximate the spoofing diversity observed in the evaluation set. By aligning the training conditions more closely with the challenges of the test environment, this augmented dataset provides a stronger foundation for training SASV systems that are more robust and generalizable.

3.2 Implementation details

For the ASV task, we adopt the Wespeaker [21] framework for implementation. Each model is trained from scratch using the bonafide labeled data from the training set of the VLSP 2025 VSASV dataset. Each audio is randomly chunked into a 3.2-second segment and then converted into log Mel-filterbank features with 80 Mel bins. The final input feature contains 200 frames, with a frame length of 25 ms and a frame shift of 10 ms. All models are optimized with the Sub-center ArcFace [22] loss function, which improves intra-class compactness and inter-class separation. The architectures and training setups of the chosen ASV models are strictly preserved to maintain their proven effectiveness.

For the CM task, we implement our system based on the XLSR-Conformer-TCM [23] repository. The CM model is also trained from scratch using the entire training set, together with the augmentation strategies described in Section 3.1, and optimized with a binary cross-entropy loss to classify between the two classes. We randomly select 10% of the data using stratified sampling for validation. In addition, several modifications are introduced into the baseline to improve the training process. In the original setup, the model outputs two logits corresponding to bonafide and spoof, with the latter directly taken as the final CM score. We instead apply a softmax to obtain the spoof probability, which yields normalized scores in $[0, 1]$. Furthermore, the initial cropping is replaced with random chunking to generate fixed-length segments for efficient mini-batch learning. At inference, the entire audio is processed with a batch size of one to leverage richer temporal information and improve detection performance. The effectiveness of these enhancements is demonstrated through empirical evaluation, reported in Table 3.

In the score fusion approach, we first compute the cosine similarity between the enrollment and test utterances using the ASV embeddings, followed by a sigmoid transformation to normalize the score into $[0, 1]$, which is more appropriate than softmax since the ASV sub-system produces a single similarity score rather than multiple competing classes. The normalized ASV score is then combined with the CM score using a weighted summation, with the weight optimized via grid search to minimize EER on the public test set.

For the embedding fusion approach, we construct training pairs of enrollment and test utterances, concatenating the ASV and CM embeddings as inputs to the DNN classifier described in Section 2.4. The classifier is trained with cross-entropy loss to produce the final spoofing-aware verification score. To ensure generalization, no speaker meta-information is used during either training or evaluation.

Table 2: EER (%) of ASV models on the public test set.

Model	**EER(%)**
Baseline	33.63
ECAPA-TDNN	32.17
MFA-TDNN	30.74
Eres2NetV2	**30.00**

3.3 Results

This section presents the performance of individual sub-systems as well as the complete SASV system, with evaluations primarily conducted on the public test set. For fair comparison and analysis, we also refer the official baseline results for each sub-system and the SASV system, provided by the organizers. The best-performing SASV configurations are further validated on the private test set.

3.3.1 ASV results

The performance of the ASV sub-system on the public test set is summarized in Table 2. The official baseline system achieves an EER of 33.63%. Among our three evaluated models, ERes2NetV2 obtains the best EER of 30.00%, followed closely by MFA-TDNN and ECAPA-TDNN. While all of our models outperform the baseline, the improvements remain relatively modest. Overall, the obtained EER values are still considerably higher than typical ASV benchmarks. We attribute this degradation mainly to two factors: (i) the limited amount of bonafide training data provided by the organizers, and (ii) the inability to fully address the noise present in the training set. In addition, we observed partial label inconsistencies in the public test set,

Table 3: EER (%) of CM models on the public test set. The last-2 models ([*i]) are used for SASV fusion.

Model	Data	EER (%)
Baseline	Original	10.11
XLSR Conformer + TCM	Original	17.04
→ w/ output softmax transform	Original	13.99
XLSR Conformer + TCM	Original + Adversarial	8.09
→ w/ random 4s chunking	Original + Adversarial + TTS	2.95
XLSR Conformer + TCM [*1]	**Original + Adversarial + TTS + VC**	**1.01**
→ w/ full-audio inference **[*2]**	Original + Adversarial + TTS + VC	1.15

Table 4: SASV fusion results using 2 different methods. EER(%) is reported on the public test set, while the last model is evaluated on both public and private test sets. The notation [*i] corresponds to model [*i] as presented in Table 3

ASV	CM	Fusion strategy	Public test	Private test
Baseline	Baseline	-	24.21	31.60
Eres2NetV2	XLSR Conformer + TCM **[*2]**	Embedding	20.60	-
Eres2NetV2	XLSR Conformer + TCM **[*1]**	Score	20.04	-
Eres2NetV2	**XLSR Conformer + TCM [*2]**	**Score**	**19.84**	**17.78**

where some utterances from the same speaker were incorrectly assigned as "non-target" and the reverse mislabeling also occurred, which further hindered performance evaluation. We also experimented with two common strategies in speaker verification, namely adaptive score normalization (AS-Norm) [24] and large-margin fine-tuning [25]. However, neither approach resulted in any performance improvements. Addressing these challenges remains an important direction for future work.

3.3.2 CM results

After reproducing the XLSR-Conformer + TCM from the repository and training it on the original data, our reproduction yielded an EER of 17.04%, which is worse than the official baseline EER of 10.11%. We then progressively incorporated the improvements described in Section 3.2, each of which contributed to noticeable performance gains. In particular, the data augmentation strategies outlined in Section 3.1 played a crucial role in achieving the final performance.

As shown in Table 3, our best model reached an EER of 1.01% on the public test set - a substantial improvement both over the official baseline (90.00% relative reduction) and over our initial reproduced model (94.07% relative reduction). These results confirm the effectiveness of our proposed enhancements and the improved generalization ability of the CM sub-system. Furthermore, the two best-performing models in Table 3 were selected as candidates for integration into the subsequent SASV fusion system.

3.3.3 SASV results

We evaluate our complete SASV system using two fusion strategies: score fusion and embedding fusion. The ASV sub-system is based on the best-performing model, ERes2NetV2, while the CM sub-system is represented by the two top-performing models: (i) XLSR-Conformer with inference on the entire utterance, and (ii) XLSR-Conformer with inference on a random segment. The overall results are summarized in Table 4. For reference, the official baseline system achieves an EER of 24.21% on the public test set and 31.60% on the private test set.

Score fusion

The optimal weight of this strategy is determined via grid search on the public test set, leveraging the availability of labels. Specifically, we represent the fusion weights as w for the ASV sub-system and $(1 - w)$ for the CM sub-system. As shown in Table 4, the best performance is achieved by combining ERes2NetV2 with the XLSR-Conformer (full-utterance inference), yielding an EER of 19.84% on the public test set. The optimal weight obtained is $w = 0.1$, indicating that the fused score relies much more heavily on the CM output. This corresponds to a relative reduction of 18.05% compared to the baseline. Consequently, the CM sub-system contributes more heavily to the final fused score,

while the ASV sub-system still plays a supporting role in the overall decision.

Embedding fusion

For this approach, only the XLSR-Conformer with full-utterance inference was considered for the CM sub-system. It achieved an EER of 20.60% on the public test set. While this result indicates that the DNN can leverage complementary information from both sub-systems, it remains less effective than score fusion, suggesting that the latter provides a more effective integration.

Private test evaluation

Due to the submission limit, only the best-performing system on the public test set was selected—score fusion of ERes2NetV2 and XLSR-Conformer with full-utterance inference.The optimal weight was initially set to $w = 0.1$ as tuned during the training phase and was later refined to $w = 0.11$ on the private test set. With this configuration, the system achieved an EER of 17.78% on the private test set, compared to the official baseline's 31.60%. This corresponds to a relative reduction of 43.73%, which is substantially superior to the baseline and even more pronounced than the improvement observed on the public test set. A plausible explanation is that the baseline system fails to generalize to novel attack types present in the private test set, whereas our augmented data endows the model with stronger robustness against diverse spoofing strategies. This highlights the effectiveness of our augmentation pipeline in handling real-world and previously unseen attacks, thereby yielding superior generalization in practical SASV applications.

4 Conclusion

In this work, we presented our system for the VLSP 2025 VSASV challenge. Our experiments reveal that the CM sub-system, based on XLSR-Conformer-TCM with extensive spoof data augmentation, substantially enhances overall performance, while the ASV sub-system remains constrained by the limited and noisy bonafide training data provided.

Among the evaluated strategies, score fusion proved to be the most effective, outperforming embedding fusion and yielding the best overall results. Specifically, our system achieved an EER of 19.84% on the public test set and 17.78% on the private test set, corresponding to relative reductions of 18.05% and 43.73% compared to the official baseline, respectively.

These results highlight the crucial role of robust spoof detection in SASV systems and suggest that future improvements may come from two directions: (i) developing more data-efficient training strategies for ASV under limited resources, and (ii) exploring advanced fusion mechanisms to better exploit the complementary strengths of ASV and CM sub-systems.

References

[1] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7):1–38, February 2025. ISSN 1557-7341. doi: 10.1145/3714458.

[2] Lam Pham, Phat Lam, Dat Tran, Hieu Tang, Tin Nguyen, Alexander Schindler, Florian Skopik, Alexander Polonsky, and Canh Vu. A comprehensive survey with critical analysis for deepfake speech detection, 2025.

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*. ISCA, October 2020. doi: 10.21437/Interspeech.2020-2650.

[4] Tianchi Liu, Rohan Kumar Das, Kong Aik Lee, and Haizhou Li. Mfa: Tdnn with multi-scale frequency-channel attention for text-independent speaker verification with short utterances. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7517–7521, 2022. doi: 10.1109/ICASSP43922.2022.9747021.

[5] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, February 2021. ISSN 1939-3539. doi: 10.1109/tpami.2019.2938758.

[6] Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, Shiliang Zhang, and Junjie Li. Eres2netv2: Boosting short-duration speaker verification performance with computational efficiency, 2024.

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *arXiv preprint arXiv:2110.01200*, 2021.

[8] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *Proc. ICASSP*, pages 6369–6373, 2021.

[9] Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. Temporal-channel modeling in multi-head self-attention for synthetic speech detection. In *Interspeech 2024*, pages 537–541. ISCA, September 2024. doi: 10.21437/Interspeech.2024-659.

[10] 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*, pages 2893–2897, 2022. doi: {10.21437/Interspeech.2022-11270}.

[11] Cunhang Fan, Bin Liu, Jianhua Tao, Jiangyan Yi, Zhengqi Wen, and Leichao Song. Deep time delay neural network for speech enhancement with full data learning, 2020.

[12] Kaavya Sriskandaraja, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah. Investigation of sub-band discriminative information between spoofed and genuine speech. In *Interspeech 2016*, pages 1710–1714, 2016. doi: 10.21437/Interspeech.2016-844.

[13] Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas Evans, and Massimiliano Todisco. An explainability study of the constant q cepstral coefficient spoofing countermeasure for automatic speaker verification, 2020.

[14] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.

[15] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching, 2024.

[16] Khanh Le, Tuan Vu Ho, Dung Tran, and Duc Thanh Chau. Chunkformer: Masked chunking conformer for long-form speech transcription. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888640.

[17] Vu Hoang. Vsasv: a vietnamese dataset for spoofing-aware speaker verification. In *Interspeech 2024*, 2024.

[18] Mengao Zhang, Ke Xu, Hao Li, Lei Wang, Chengfang Fang, and Jie Shi. Doubledeceiver: Deceiving the speaker verification system protected by spoofing countermeasures. pages 4014–4018, August 2023. doi: 10.21437/Interspeech.2023-371.

[19] Hendrik Schröter, Tobias Rosenkranz, Alberto N. Escalante-B., and Andreas Maier. DeepFilterNet: Perceptually motivated real-time speech enhancement. In *INTERSPEECH*, 2023.

[20] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[21] Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[22] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[23] Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. Temporal-channel modeling in multi-head self-attention for synthetic speech detection. *arXiv preprint arXiv:2406.17376*, 2024.

[24] Jeong-Hwan Choi, Ju-Seok Seong, Ye-Rin Jeoung, and Joon-Hyuk Chang. Trainable adaptive score normalization for automatic speaker verification. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890182.

[25] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck. The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5814–5818, 2021. doi: 10.1109/ICASSP39728.2021.9414600.