Interspeech 2024
1-5 September 2024, Kos, Greece

VSASV: a Vietnamese Dataset for Spoofing-Aware Speaker Verification

Vu Hoang[*], *Viet Thanh Pham*[*], *Hoa Nguyen Xuan*[*], *Nhi Pham, Phuong Dat, Thi Thu Trang Nguyen*[†]

Hanoi University of Science and Technology, Vietnam

{longvu200502, thanh.pv.ds, thaihoahochust, nhi.phamt2002, phuongtuandat2915}@gmail.com, trangntt@soict.hust.edu.vn

Abstract

Recent research in improving speaker verification systems to detect spoofed speech has seen a concentrated focus on English language, while the performance of such systems in other languages remains unexplored. This paper introduces the VSASV dataset for Spoofing-Aware Speaker Verification (SASV) in Vietnamese language. The dataset comprises over 174,000 spoofed utterances and 164,000 authentic utterances from 1,382 speakers, which were generated with the latest spoofing techniques to encourage the development of SASV systems in this language. We also provide experimental results on the efficacy of the different state-of-the-art anti-spoofing systems on Vietnamese language.

Index Terms: speaker verification, spoofing-aware, speaker recognition

1. Introduction

Automatic speaker verification (ASV) is the task of verifying the identity of an individual based on their speech segments. ASV systems have been seen as a cost-effective and convenient method for biometric identification of a person. ASV has extensive applications in voice-activated services, from interactive voice-based assistants to authentication systems. Nevertheless, ASV has been proved to be vulnerable to spoofing attacks ([1]).

Previous works ([2], [3]) suggest various effective typical attack methods, featuring: voice conversion, text-to-speech (TTS), and replay attacks. While replay attacks rely on playing back or mimicking the original voice, the other advanced methods aim to create synthetic speech that is indistinguishable from human perception.

Several datasets have been created to facilitate research and training systems for detecting voice spoofing. ASVspoof 2019 [4] introduces a new dataset of bona fide and spoofed speech signals, including synthetic speech and converted voice signals generated with the state-of-the-art technologies to establish common platforms for the comparative evaluation of different spoofing solutions. Along with ASVspoof 2019, VCTK, Spoofing and Anti-Spoofing (SAS) are publicly available datasets that mainly focus on English language. However, there has been no datasets for the purposes of training anti-spoofing systems in Vietnamese language. Therefore, the performance of models on this language remains unexplored, since the authors of HABLA dataset [5] suggest that the variability in pronunciation and accent can severely degrade the generalisation of such models on different languages. The recent advances in spoofing techniques based on generative models have also posed an urgent need of a comprehensive dataset in Vietnamese language that is capable of providing a platform for training and evaluating the effectiveness of existing anti-spoofing models.

This paper introduces the construction of the first Vietnamese dataset for training and evaluating purposes of anti-spoofing and voice recognition models, Vietnamese Spoofing-aware Speaker Verification (VSASV) dataset. The dataset contains about 340,000 utterances from nearly 1,400 people, covering comprehensively various genders and dialects in Vietnam. The dataset also considers the latest, state-of-the-art technologies in generating spoofed signals to encourage the development of more secure anti-spoofing and ASV systems. Finally, we present the efficacy of current state-of-the-art anti-spoofing systems in our new dataset.

The subsequent sections in this paper are organized as follows: Section 2 describes the construction of VSASV dataset, Section 3 discusses the statistics of the dataset along with the division strategy for further experimental evaluations in Section 4. Section 5 concludes our contributions in this paper.

2. Data Construction

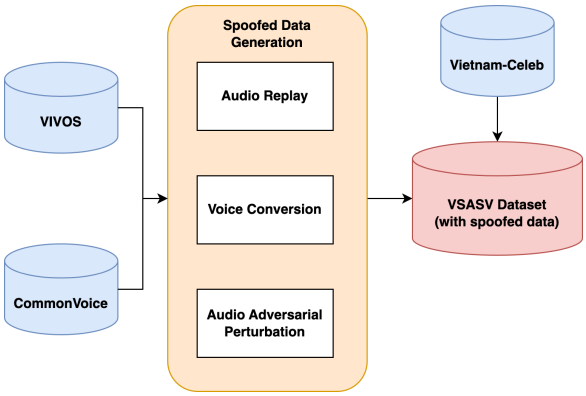

Figure 1: *The construction pipeline of VSASV dataset.*

The VSASV dataset construction pipeline is illustrated in Figure 1, utilizing three publicly available datasets of Vietnamese utterances: (1) Vietnam-Celeb [6] - our published dataset of 1,000 speakers covering different dialects and genders of Vietnamese voices, (2) CommonVoice - a multilingual speech corpus, with 6 hours of Vietnamese language from 252 different speakers, and (3) VIVOS - a free Vietnamese speech corpus, consisting of 15 hours of utterances from 65 different

[*]Authors equally contributed to this work.
[†]Corresponding author.

10.21437/Interspeech.2024-1972

speakers. The pipeline comprises the spoofed data generation step from the bona fide utterances provided by CommonVoice and VIVOS. The generation process applies three spoofing techniques: audio replay, voice conversion, and audio adversarial perturbation. The final VSASV dataset includes bona fide utterances from Vietnam-Celeb, CommonVoice, VIVOS, and the generated spoofed utterances from the two latter corpora.

2.1. Spoofed data generation

Since speech segments in Vietnam-Celeb corpus have significant noises (background noise, background music, etc.), these are not efficient for spoofing attack setups, the reason for which will be further explained in the subsequent description of each spoofing technique. As a result, we leverage the speech segments in Vietnamese language from VIVOS and CommonVoice for the spoofing examples generation.

To reflect the best as possible the practical scenario, and to encourage the development of robust anti-spoofing systems, we consider the latest state-of-the-art spoofing attack technologies, including audio replay, voice conversion, and adversarial attack to generate spoofed utterances for VSASV dataset.

2.1.1. Audio Replay

Replay attacks are the most straightforward method to implement, by recording a bona fide utterance. As the spoofed utterances are simply recordings of bona fide speech, replay attacks can be highly effective in deceiving ASV systems and may cause a higher false-acceptance rate than other complicated spoofing techniques like voice conversion or speech synthesis, as discussed in [7].

Replay attacks are conducted by replaying an original bona fide utterance. We used various personal laptops and smartphone devices with reasonable quality in-built speakers as replay devices, and carried out the replay in an office room environment. Owing to limitations in recording devices, we set up the input and output devices to be the same device for one replay session. The original bona fide utterances are Vietnamese segments from two publicly available speech corpora: CommonVoice[8] and VIVOS [9]. Since the utterances from Vietnam-Celeb corpus are collected on the social media with significant background noises (as shown in Figure 2), the replayed audios can be easily distinguished, making replay attacks for audios from Vietnam-Celeb less effective. After replaying on all of the Vietnamese utterances from VIVOS and CommonVoice, we filter out those with significant noises in the audio (based on Signal-to-noise ratio) to collect over 55,000 valid replayed utterances.

2.1.2. Voice Conversion

Voice conversion (VC) is a spoofing attack against automatic speaker verification. This method utilizes a natural voice from the attacker, then converts into a speech of the target. VC does not depend on a text input, but operates directly on speech inputs. Active research in deep learning models marked a milestone of advanced voice conversion techniques, tremendously improved the voice quality and similarity to target speaker ([10]).

This paper utilizes Retrieval-based Voice Conversion (RVC), a technique that uses a corpus of pre-recorded speech to create a synthetic copy of a person's voice. RVC works by finding the pre-recorded speech that is most similar to the input speech. Then the input speech is converted so as to sound like the pre-recorded speech. We used RVC Project, a public voice conversion toolkit that allows fine-tuning on new voices. RVC uses a pre-trained UVR5 model to quickly separate vocals and instruments, along with a pre-trained HuBERT for speech representation.

The RVC toolkit was used to fine-tune on Vietnamese speeches in CommonVoice and VIVOS corpora, since the Vietnam-Celeb corpus contains significant background noises. The input to the models require one pair of speeches from two different speakers. The speakers are chosen so that the cosine similarity between two embeddings are greater than a threshold, therefore the synthetic speech is more indistinguishable. Afterwards, the synthetic speeches undergo a filtering process, where only the synthetic speeches with high cosine similarity to the target speaker's speech remain.

2.1.3. Adversarial Attack Simulation

Related works in mitigating the threats of spoofing attacks introduced joint systems of anti-spoofing model and countermeasures model as a state-of-the-art solution ([11], [12]). However, most research only consider a white-box scenario rather than a black-box situation in real-world setup, where the identity of the sub-systems are unknown to the attackers. DoubleDeceiver [13] proposed an adversarial attack technique utilizing synthetic voice generated from a Text-to-Speech (TTS) system to mimic the target speaker by adding adversarial perturbation, achieving a successful attack rate (SAR) as high as 98.3%. To encourage the development of more robust anti-spoofing models, we employed DoubleDeceiver to generate challenging adversarial examples in our VSASV dataset.

DoubleDeceiver works by feeding a synthesized voice by the TTS system into the surrogate anti-spoofing and ASV model. The gradient of both losses of the models is then combined to perform gradient-based adversarial attack method to get the adversarial perturbation. The candidates for surrogate ASV model include: ECAPA-TDNN [14] and ResNet34 [15]; while AASIST [16] and S^2pecNet [17] are used as a surrogate anti-spoofing model.

3. The VSASV Dataset

3.1. Overall Statistics

Table 1: *The number of speakers and hours for each type of utterance*

Utterance type	# of Speakers	# of Hours
Bonafide	1,382	356.36
Replay	46	30.04
Voice conversion (VC)	147	119.44
Adversarial perturbation (AP)	317	35.18

The number of speakers and number of hours of each utterance type is shown in Table 1. The VSASV dataset contains 541.02 hours of short utterances from 1,382 Vietnamese speakers. The bona fide audio samples originate from pre-existing Vietnamese datasets, namely Vietnam-Celeb, CommonVoice and VIVOS.

The noise distribution in each subset of the VSASV dataset is computed by evaluating the Signal-to-Noise Ratio (SNR) [18] value. As illustrated in Figure 2, the majority of utterances in the Vietnam-Celeb subset falls within the range of 0 to 20

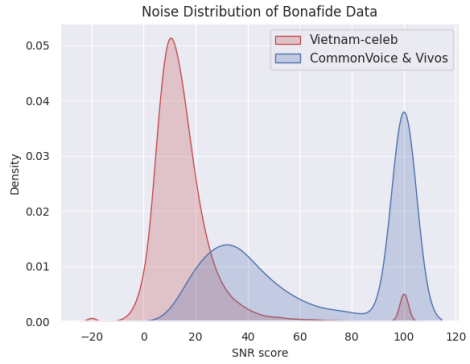

Figure 2: *Noise distribution of the Bona fide data.*

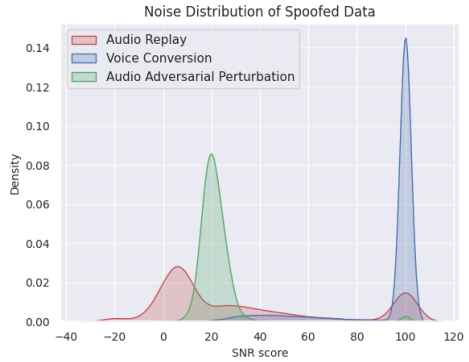

Figure 3: *Noise distribution of the Spoofed data.*

dB, meaning that the dataset contains a diverse set of recordings with varying levels of background noise. These SNR values capture a wide range of real-world scenarios, reflecting the acoustic environments in which speakers typically operate. We observed that a small portion of the bona fide set, consisting of utterances derived from CommonVoice and VIVOS publicly available speech corpora, exhibits a higher concentration of high SNR values. The presence of these clean recordings provided a reliable source for the spoofing examples generation.

Three types of spoofed examples were generated using the speech segments from the two publicly available VIVOS and CommonVoice dataset. We employed Voice Conversion technique, where a natural voice is transformed to mimic a target speaker's voice. A subsequent filtering process described in Section 2.1.2 is applied to mitigate excessive noise in the synthetic output. Another spoofing approach involves incorporating adversarial perturbations directly into the speech corpus. This manipulation has been observed to increase background disruptions as illustrated in Figure 3. We also implement the Audio Replay attack, where pre-recorded speech is simply replayed. This attack, while replicating the overall speech patterns of the VIVOS and CommonVoice corpora, exhibits a rise in noise level due to the differences in recording devices used during the original capture.

3.2. Train-Test Split

The SASV 2022 Challenge [12] observed that the combination of ECAPA-TDNN [14] and AASIST [16] model has demonstrated effectiveness in detecting spoof audio generated from voice conversion and audio replay techniques. We constructed the training dataset using bona fide utterances and two spoofing techniques, namely voice conversion and replay. For the testing dataset, we partitioned into various small scenarios, each representing a combination of distinct spoofing techniques: audio replay (Replay), voice conversion (VC), and adversarial perturbation (AP). This segmentation allowed us to assess the model's effectiveness against each type of attack. Table 2 respectively indicates the statistics of training and testing set of VSASV dataset.

Table 2: *The number of utterances of each speech type in each corpus*

Utterance type	Training	Testing
Bonafide	132,424	31,950
Voice conversion (VC)	37,292	63,272
Replay	23,320	34,251
Adversarial perturbation (AP)	N/A	16,731

4. Experiments

4.1. Model Architecture

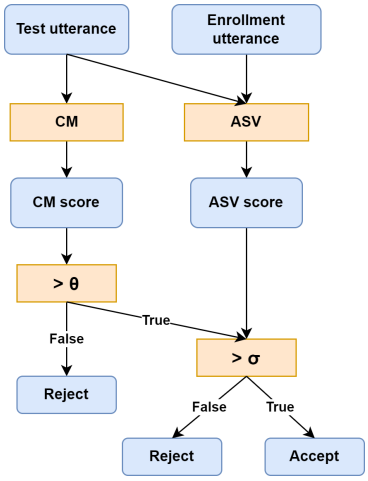

Figure 4: *Model Architecture of SASV System.*

The model architecture is illustrated in Figure 4. This system comprises two standalone modules: automatic speaker verification (ASV) and countermeasure (CM) sub-system. If only the ASV is used, it will extract ASV embeddings from the test and enrollment utterances to calculate their similarity score. If we combine this sub-system with the countermeasure, CM score will be extracted from the test utterance using spoofing detection model. If this score is greater than a given threshold, this utterance will be fed to ASV system along with its enrollment, otherwise it will be considered as spoofed and then rejected.

4.1.1. Automatic Speaker Verification

Automatic speaker verification systems determine whether an input utterance belongs to a given speaker. ECAPA-TDNN has shown efficiency in speaker verification. It builds upon the suc-

Table 3: *Results of different baselines to different kinds of attack using EER. (a) ECAPA-TDNN pretrained on VoxCeleb, (b) ECAPA-TDNN trained on bonafide utterances only, (c) ECAPA-TDNN in (b) combined with AASIST trained on replay and voice conversion utterrances*

Test case	ECAPA-TDNN (a)	ECAPA-TDNN (b)	ECAPA-TDNN & AASIST (c)
Bonafide data only	2.29	1.15	1.15
Bonafide + AP	12.87	8.96	10.31
Bonafide + VC	2.60	1.07	2.24
Bonafide + VC + AP	10.73	7.87	9.09
Bonafide + Replay	20.12	15.69	2.51
Bonafide + Replay + VC	17.08	13.53	2.41
Bonafide + Replay + AP	25.44	17.46	9.23
Bonafide + Replay + AP + VC	22.03	15.58	8.27

cess of Time Delay Neural Networks (TDNNs) by incorporating several enhancements. ECAPA-TDNN focuses on specific regions within the speech feature channels, allowing it to extract speaker-discriminative information more effectively.

4.1.2. Spoofing detection

Spoofing detection task is to determine whether a given speech utterance is genuine (bona fide) or spoofed. To encounter spoofed utterances lying among the bona fide ones, a countermeasure system is necessary to detect these attacks. AASIST is an efficient end-to-end spoofing countermeasure system. It extracts features from audio data to specify if a given audio is bona fide or spoofed.

4.2. Evaluation metrics

SASV system performance is measured by classical Equal Error Rate (EER). We define *target* and *non-target* as in [19]. Besides the reliability of the ASV system, the false acceptance of the countermeasure towards spoofed utterance can cause to increase EER.

4.3. Experimental results

Three SASV systems are employed for the test cases. The first system (a) comprises an ECAPA-TDNN pre-trained on VoxCeleb2 of mainly English utterances, and an AASIST subsystem trained on spoofed utterances following the architecture in Figure 4; (b) follows the architecture in Figure 4, consisting of an ECAPA-TDNN trained on bona fide utterances of VSASV dataset, and an AASIST sub-system trained on spoofed utterances following; (c) is the fused system of ECAPA-TDNN from (b) with an AASIST trained on spoofed utterances. The models were trained on a single NVIDIA Tesla V100 32GB GPU with the training parameters such as learning rate, batch size implemented as the original papers of each model.

Based on the experimental results given in the Table 3, we first observed a significant decrease of 158% in average in the verification performance for all test cases when applying the SASV systems to a language different from the trained language. This shows a lack of multi-language capability of SASV systems.

In terms of the performance for each of the spoofing attack techniques, it can be seen that generating replayed utterances from clean audios, in an office environment helps deceive standalone ASV systems effectively despite its simple implementations setup, as explained in [7] and [20]. Audio replay caused a relative increase of 1100% in EER for SASV systems in average. Moreover, adversarial perturbation also shows significant threats to ASV systems, with a relative increase of 780% and 560% in EER when compared to bona fide test only for systems trained on Vietnam-Celeb and Vox-Celeb2, respectively. Its participant also shows efficiency in deceiving ASV systems when combined with other spoofing techniques (replay and voice conversion).

The improvement in the performance of the ECAPA-TDNN (b) system on the test case of VC utterances may suggest that the countermeasure sub-sytem managed to distinguish the voices generated by VC technique well. Moreover, the spoofed utterances generated by VC observe a high concentration in the high values range of SNR (as shown in Figure 3), which may lead to a significant resemblance between the utterances used for training and testing. The presence of a countermeasure sub-system managed to improve the performance of the ASV system by 400% in average, for 4 cases of spoofing attacks: Replay, Replay + VC, Replay + AP, and Replay + VC + AP. However, since the authors of DoubleDeceiver discussed that adversarial perturbation only targets at attacking the combination of ASV and countermeasure sub-system, we observed that the standalone ASV system outperforms the combination in the cases relating to adversarial perturbation alone.

5. Conclusions

This paper introduced VSASV, the first dataset of Vietnamese language for voice anti-spoofing, publicly available under a GitHub repository[1]. The dataset contains over 164,000 bona fide utterances of 1,382 speakers from Vietnam-Celeb, VIVOS, and CommonVoice. The spoofed utterances are generated from the two latter datasets following the current state-of-the-art spoofing technologies, totalling 174,000 utterances, to encourage the training of robust SASV systems for Vietnamese voices.

Various experiments were carried out to show how each of the applied spoofing technology affect the performance of ASV systems. We observed significant improvements in detecting spoofed utterances by employing a state-of-the-art countermeasure subsystem fused with ASV. Meanwhile, there was a tremendous degradation in the performance when evaluating on a language different from the one that ASV model was trained on. The experimental results also showed that our generated spoofed utterances posed significant threats to current best ASV systems, since they were generated from large-scale datasets using the latest, state-of-the-art spoofing techniques. Therefore, we believe that the proposed VSASV dataset can be a valuable resource for improving the capability to detect spoofed voices in Vietnamese language of SASV systems.

[1] https://github.com/hustep-lab/VSASV-Dataset

6. References

[1] P. Gupta, H. A. Patil, and R. C. Guido, "Vulnerability issues in automatic speaker verification (asv) systems," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, pp. 1–14, 2024.

[2] H. Dinkel, Y. Qian, and K. Yu, "Investigating raw wave deep neural networks for end-to-end speaker spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2002–2014, 2018.

[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

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

[5] P. A. Tamayo Flórez, R. Manrique, and B. Pereira Nunes, "HABLA: A Dataset of Latin American Spanish Accents for Voice Anti-spoofing," in *Proc. INTERSPEECH 2023*, 2023, pp. 1963–1967.

[6] H. L. V. N. T. T. T. Pham Viet Thanh, Nguyen Xuan Thai Hoa, "Vietnam-celeb: a large-scale dataset for vietnamese speaker recognition," 2023.

[7] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, 2014, pp. 1–5.

[8] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[9] H.-T. Luong and H.-Q. Vu, "A non-expert kaldi recipe for vietnamese speech recognition system," in *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)*, 2016, pp. 51–55.

[10] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[11] W. Zhang, S. Zhao, L. Liu, J. Li, X. Cheng, T. F. Zheng, and X. Hu, "Attack on practical speaker verification system using universal adversarial perturbations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2575–2579.

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

[13] M. Zhang, K. Xu, H. Li, L. Wang, C. Fang, and J. Shi, "Doubledeceiver: Deceiving the speaker verification system protected by spoofing countermeasures."

[14] K. D. Brecht Desplanques, Jenthe Thienpondt, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," 2020.

[15] B. Koonce and B. Koonce, "Resnet 34," *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pp. 51–61, 2021.

[16] H. T. H.-j. S. J. S. C. B.-J. L. H.-J. Y. N. E. Jee-weon Jung, Hee-Soo Heo, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," 2021.

[17] P. Wen, K. Hu, W. Yue, S. Zhang, W. Zhou, and Z. Wang, "Robust Audio Anti-Spoofing with Fusion-Reconstruction Learning on Multi-Order Spectrograms," in *Proc. INTERSPEECH 2023*, 2023, pp. 271–275.

[18] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," 2008.

[19] H.-j. S. H.-S. H. B.-J. L. S.-W. C. H.-J. Y. N. E. T. K. Jee-weon Jung, Hemlata Tak, "Sasv 2022: The first spoofing-aware speaker verification challenge," 2022.

[20] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *2014 International conference of the biometrics special interest group (BIOSIG)*. IEEE, 2014, pp. 1–6.