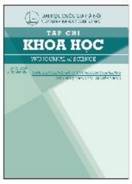
VNU Journal of Science:
Computer Science and Communication Engineering

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



Original Article

SV - VLSP 2021: Combine Attentive Statistical Pooling-based Xvector and Pretrained ECAPA-TDNN for Vietnamese Text-Independent Speaker Verification

Ta Bao Thang[1,2], Huynh Thi Thanh Binh[2,*]

[1]Viettel Cyberspace Center, Ton That Thuyet, Cau Giay, Hanoi, Vietnam
[2]Hanoi University of Science and Technology, 1 Dai Co Viet, Hai Ba Trung, Hanoi, Vietnam

Received 27 December 2021
Revised 5 April 2022; Accepted 5 May 2022

Abstract: Recently, Xvectors and ECAPA-TDNN have been considered state-of-the-art models in designing speaker verification systems. This paper proposes a novel approach that combines Attentive statistic pooling-based Xvector and pre-trained ECAPA-TDNN for Vietnamese speaker verification. Experiments are conducted on various recent Vietnamese speech datasets. The results portrayed that our proposed combination outperformed all constitutive models with 4% to 37% relative EER improvement and ranked second place in Task 2 of the 2021 VLSP Speaker Verification competition.

Keywords: Speaker Verification, Xvector, Attentive Statistical Pooling

1. Introduction

Speaker Verification (SV) is an important biometric problem attracting significant attention from the research community and industry due to its urgent applications in practice. For example, investigators collected audio samples of a perpetrator at the scene. The investigators want to compare audio samples of each suspect and the obtained audio to verify whether there is a perpetrator or not. Besides, users can use their voice to open, verify, or make decisions in payment applications instead of using traditional methods. The SV system includes two main steps: registration and verification. First, the user registers one or several voice samples. Next, the system extracts the speaker's acoustic features and saves them in the database. These features are considered as the "voice signature" of each person. Then, in the verification process, the user provides his/her identity and voice sample. The system extracts the feature, compares it with the stored acoustic feature, and gives a successful or failed validation message. SV systems can be divided into two types:

[*] Corresponding author.
E-mail address: binhht@soict.hust.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.320

• Text-Independent Speaker Verification (TI-SV) is based only on acoustic features, independent of the content.

• Text-Dependent Speaker Verification (TD-SV) depends on the content of the speech. This content (password) is fixed and needs to be registered first.

Compared with TD-SV, TI-SV is more flexible because it does not force users to say a fixed content but can say whatever they want. Recently, many approaches have been proposed for TI-SV. However, there is little effort put into the Vietnamese language. This paper presents a combination of two recent state-of-the-art models: Attentive statistical pooling-based Xvector [1] and ECAPA-TDNN [2] for Vietnamese Text-Independent Speaker Verification. Our proposal outperformed all constitutive models and achieved second place in Task 2 of the Speaker Verification Competition held at the eighth workshop on Vietnamese Language and Speech Processing (VLSP 2021)[3]. The rest of this paper is organized as follows. Section 2 introduces several recent approaches for the TI-SV problem, while the proposed model is described in Section 3. Section 4 presents the experimental scenarios, settings, and obtained results. Discussions and future works are drawn in Section 5.

2. Related Works

The past decade has witnessed the success of Deep Neural Network (DNN) models for speaker verification systems. Ghalehjegh [4] proposed a DNN model using the i-vector feature. Snyder [5] replaced i-vector with features extracted from DNN models, allowing to distinguish speakers by non-fixed length signal inputs. The results showed that the proposal is better than i-vector on short input and competes in long input. In that same year, Zhang [6] proposed an end-to-end model using Inception Net. One year after, Torfi [7] proposed a 3D Convolution Neural Network.

Among DNN-based models, Xvector [8] and its variants [9, 10] have been state-of-the-art methods for the SV problem. The input of these models often is Mel Frequency Cepstral Coefficients (MFCCs) because it can capture essential features into a small dimension representation [1, 8, 9]. Furthermore, to adapt non-fixed length inputs, most DNNs models implemented a pooling layer to aggregate frame-level embeddings into a fixed-length embedding. In the Xvector model, statistical pooling is implemented to calculate frame-level features' mean and standard deviation. However, this method assigns equal weight to all frames, ignoring the importance of some special frames. Therefore, some recent studies have added an attention layer to pooling layers. For example, Okabe [1] weighted on the mean and standard deviation of the signal frames. These weights are learned by an attention mechanism. Zhu [10] introduced a pooling mechanism based on self-attention and multiple attention heads. However, a common limitation of these methods is that the weights of the signal frames are scalar. As a result, the elements in each frame have the same weight when calculating the mean and standard deviation, leading to missing essential features. Recently Desplanques [2] proposed an ECAPA-TDNN model using 1D Res2Net blocks with shortcut connections to improve the previous Xvector model.

However, all the models mentioned above are proposed for the English language. There are few studies for the Vietnamese language. In this paper, we proposed a hybrid model for Vietnamese verification systems. The proposed model combines features obtained from the attentive-statistical pooling based Xvector [1] and an English pre-trained ECAPA-TDNN model. The reason is that languages often have a lot in common, and a system trained to distinguish speakers in a specific language can also distinguish speakers in other languages well. Utilizing an English pre-trained model helps us achieve better performance without additional Vietnamese data. Besides, combining multiple different models can take advantage of constitutive models' mutual complementarity.

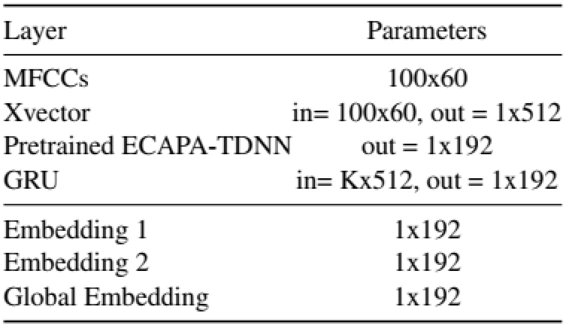
3. Proposed Model

The proposed model architecture is presented in Figure 1.

Firstly, input voice files are split into one-second signal segments. Then, MFCC features are extracted from these signal segments to build the Attentive-pooling-based Xvector classification model. The loss function used is Cross-Entropy Loss. As a result, we obtain a set of time-sequential Xvector embeddings for each voice input. Next, these embeddings are fed through the GRU network with the Generalized end-to-end loss (GE2E)[11] to obtain a more miniature representation that captures the essential characteristics in the more extended context of each speaker.

Besides, a recently proposed model - ECAPA-TDNN [2], previously trained for English and provided by Speech Brain library1, is used in this paper to extract the speaker embeddings. This approach hopes to take advantage of the power of other language models in Vietnamese. Finally, embeddings obtained from the two mentioned-above approaches are normalized and concatenated to obtain a global speaker embedding for each speech input file. The configuration of the proposed model is shown in Table 1.

Table 1. Configuration of the proposed model:



Layer	Parameters
MFCCs	100x60
Xvector	in= 100x60, out = 1x512
Pretrained ECAPA-TDNN	out = 1x192
GRU	in= Kx512, out = 1x192
Embedding 1	1x192
Embedding 2	1x192
Global Embedding	1x192

4. Evaluation

4.1. Dataset

The dataset for the Speaker Verification in this paper is taken from the Speaker Verification contest at the 8th International Workshop on Vietnamese Language and Speech Processing (VLSP 2021). The training dataset includes about 35k files of 1305 speakers. The mean length of these files was 4 seconds. Two private tests T1 and T2, provided by the contest, include 20k pairs of files.

4.2. Experimental Scenarios

We evaluate the performance of our proposed combined model and two constitutive models (Xvector, ECAPA-TDNN). All models use Adam, learning rate 3×10^{-4} and batch size 15 on a GPU device. No data preprocessing or augmentation techniques are used in this proposal.

Besides, we also verify the efficiency of the proposed model with several other models in the 2021 VLSP competition.

4.3. Experimental Criteria

The paper adopts the Equal Error Rate (EER) to evaluate the effectiveness of the proposed model. EER is the point when the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) of the verification system are equal. The meaning of this indicator is explained as follows.

The higher the FRR, the more secure the system is. However, when that happens, many legal users' validations are rejected. These users have to spend a considerable effort to get a successful message, which reduces the user experience. Therefore, the sensitivity and convenience of the system are poor.

On the contrary, if FRR is too small, the FAR is often remarkably high. As a result, the system accepts a lot of invalid user verifications. It is easy for users to authenticate successfully, but there is a very high risk in security.

The point where the FAR = FRR is called the Equal Error Rate (EER). At this point, the system balances between security and sensitivity, convenience. Therefore, EER is often used as a metric for verification systems. The smaller the EER, the better performance.

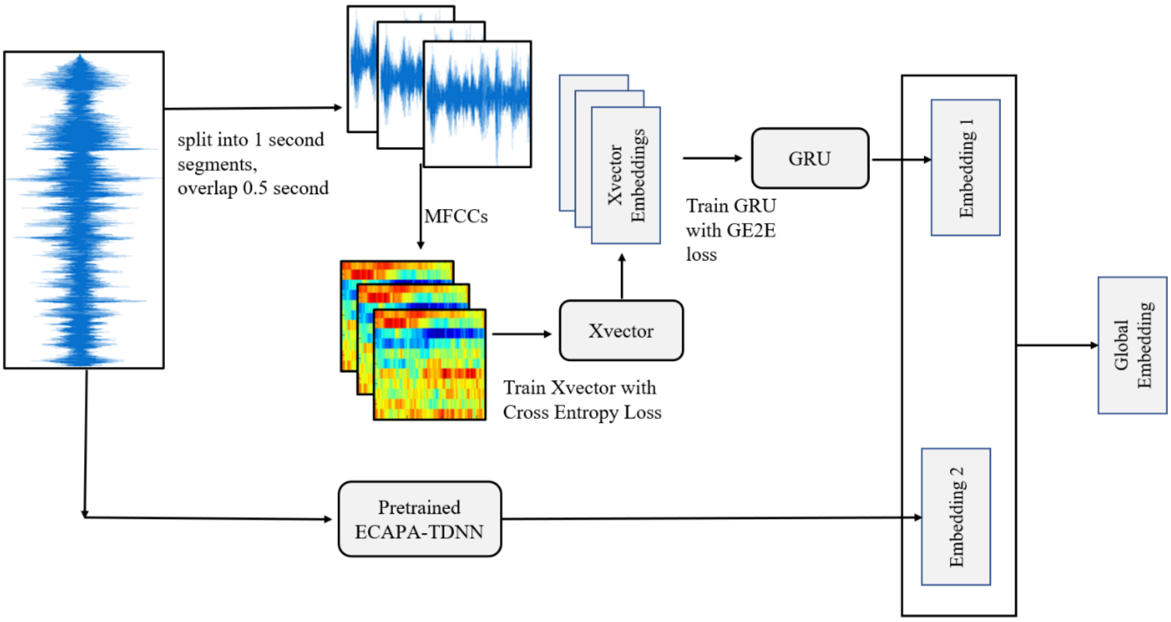


Figure 1. Proposed speaker verification model.

4.4. Experimental Results

To verify the effectiveness of the proposed combined model, we compare it with constitutive models individually, as shown in Figure 2. The obtained results showed that the proposed combined model outperformed all constitutive models on both two private datasets. Compared to ECAPA-TDNN, our proposal got 4% and 31% relative EER improvement in private test T1 and T2, respectively. Compared to Attentive-pooling-based Xvector, our combined model resulted in 36% and 37% improvement in EER values. This proved the effectiveness of our combination.

Besides, to further examine the efficacy of our proposal, we also compare it with the top-5 competitors' models in the 2021 VLSP as shown in Table 3. The results showed that our proposal achieved second place and had better improvements than three remaining competitors, about 2.2%, 10%, and 44.3%, respectively. However, our proposal under-performed by 4.5% of EER when compared with the first rank model.

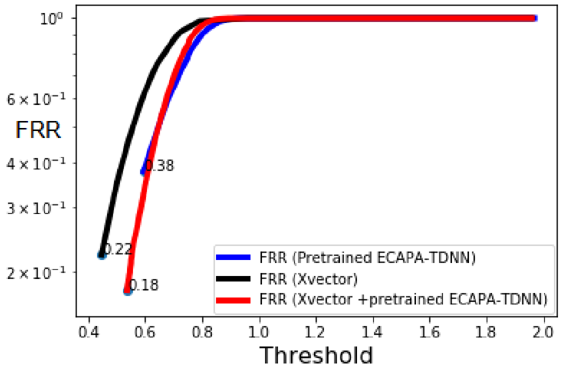


Figure 2. Our FRR when FAR < 1%.

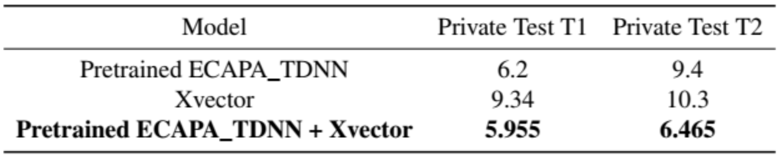
Finally, we check the practical applicability of our model. Suppose that the verification systems need a FAR of no more than 1% to meet practical security requirements. In other words, if a user is an attacker, the probability of passing through the system is less than 1%. With a defined FAR threshold, the paper checks how much the FRR of our model is.

The lower the FRR, the more convenient for the user. Consider a probability that a legal user's verification is denied is r. Let x be the number of attempts for this user to have a successful response. The expectation of x is calculated as follows:

$$E[x] = \int_{x=1}^{\infty} x \times r^{x-1} \times (1-r) dx = (1-r) \times \frac{1 - ln(r)}{ln^2(r)} \quad (1)$$

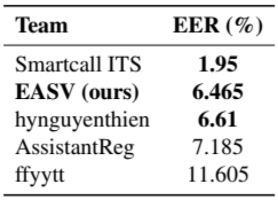
The obtained FRR values are shown in Figure 2. The finding showed that our combined model achieved the smallest FRR (18%) among all experiments. With FRR = 18%, E(x) = 0.75. It means that a legal user can be verified successfully on the first try. Overall, our system is quite convenient for users.

Table 2. Compare EER of our combined model with constitutive models.



Model	Private Test T1	Private Test T2
Pretrained ECAPA_TDNN	6.2	9.4
Xvector	9.34	10.3
Pretrained ECAPA_TDNN + Xvector	**5.955**	**6.465**

Table 3. Compare EER of our best model with top-5 competitor's models in the 2021 VLSP:



Team	EER (%)
Smartcall ITS	**1.95**
EASV (ours)	**6.465**
hynguyenthien	**6.61**
AssistantReg	7.185
ffyytt	11.605

5. Conclusion

This paper proposes a novel model combining two state-of-the-art approaches: Xvector and ECAPA-TDNN for Vietnamese Speaker Verification. Our combination outperformed all constitutive models with EER improvement of 4% to 37%. The proposed model also proved its convenience when considering high-security scenarios. Our proposal also achieved second prize in the 2021 VLSP Speaker Verification Competition. In the future, further studies need to be conducted to run the model in real-time. Besides, we can let users register more voice samples to increase model performance.

References

[1] K. Okabe, T. Koshinaka, K. Shinoda, Attentive Statistics Pooling for Deep Speaker Embedding, in: Proc. Interspeech 2018, pp. 2252–2256. doi:10.21437/Interspeech.2018-993.

[2] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, in: Proc. Interspeech 2020, 2020, pp. 3830–3834. doi:10.21437/Interspeech.2020-2650.

[3] V. T. Dat, P. V. Thanh, N. T. T. Trang, SV Challenge: Vietnamese Speaker Verification in noisy environments, VNU Journal of Science: Computer Science and Communication Engineering, VLSP 2021.

[4] S. H. Ghalehjegh, R. C. Rose, Deep Bottleneck Features For I-Vector Based Text-Independent Speaker Verification, In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015, pp. 555–560.

[5] D. Snyder, D. Garcia-Romero, D. Povey,
S. Khudanpur, Deep Neural Network Embeddings for Text-Independent Speaker Verification., in: Interspeech, 2017, pp. 999–1003.

[6] C. Zhang, K. Koishida, End-to-end Text-Independent Speaker Verification with Triplet Loss on Short Utterances, in: Interspeech, 2017, pp. 1487–1491.

[7] A. Torfi, J. Dawson, N. M. Nasrabadi, Text-independent Speaker Verification Using 3d Convolutional Neural Networks, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2018, pp. 1–6.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust Dnn Embeddings For Speaker Recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE,

2018, pp. 5329–5333.

[9] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, S. Khudanpur, Speaker Recognition for Multi-Speaker Conversations Using X-Vectors, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5796–5800.

[10] Y. Zhu, T. Ko, D. Snyder, B. Mak, D. Povey, Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification., in: Interspeech, Vol. 2018, 2018, pp. 3573–3577.

[11] L. Wan, Q. Wang, A. Papir, I. L. Moreno, Generalized End-To-End Loss for Speaker Verification, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 4879–4883.