SOURCE SPEAKER TRACING CHALLENGE

Akshet Patial

Indraprastha Institute of Information Technology
New Delhi, India
akshet23155@iiitd.ac.in

ABSTRACT

Voice conversion systems can alter audio to imitate a different speaker's voice, potentially compromising speaker verification systems. However, research on source speaker verification faces challenges due to limited data and methodological issues. This paper addresses these challenges by using an extensive converted speech database and training several baseline systems using ECAPA-TDNN [1] to enhance source speaker verification.

***Index Terms*—** Automatic Speaker Verification, source speaker verification, voice conversion

1. INTRODUCTION

Automatic Speaker Verification (ASV)[2] [3] [4] [5] systems are designed to authenticate speaker identities using voice samples. With advancements in deep neural networks, ASV [6] systems have been increasingly integrated into various applications, showcasing remarkable performance across different scenarios. But still VC Voice conversion pose threat to the safety of the users. VC technology works by converting the original speaker's voice into one that matches a target speaker's while keeping the linguistic features same. Recently, researchers and institutions have organized the Automatic Speaker Verification Spoofing and Countermeasures Challenges to develop various defense mechanisms against VC spoofing attacks. We have used the converted speech dataset to train our model to test the performance on the test set.

2. METHOD

We have used ECAPA-TDNN model for extracting the embedding of the speakers it takes raw audio as the input unlike the MFA-Conformer [7] which takes spectrogram. The ECAPA-TDNN stands for Emphasized Channel Attention, Propagation, and Aggregation-Time Delay Neural Network. Ecapa works on following layers

Squeeze-Excitation Blocks: These blocks emphasize important channels by recalibrating channel-wise feature responses consisiting of Global average pooling across time. Intermediate Feature Aggregation Layers which includes attention Mechanism. It contains Multi-Head Self-Attention, Multiple attention heads capture different aspects of the input sequence. The outputs from all heads are concatenated and transformed by a fully connected layer.

Utterance-Level Representation Layer: This layer combines the aggregated features from the previous layer to form the final utterance-level representation.

By combining these layers and components, ECAPA-TDNN effectively captures both local and global speaker-specific features, achieving state-of-the-art performance in speaker verification tasks. ECAPA-TDNN can also work with the raw audio instead converting them into Spectrogram. Given the audio in the Input Ecapa -TDNN gives embeddings which has been used to verify the speakers as per the required task.

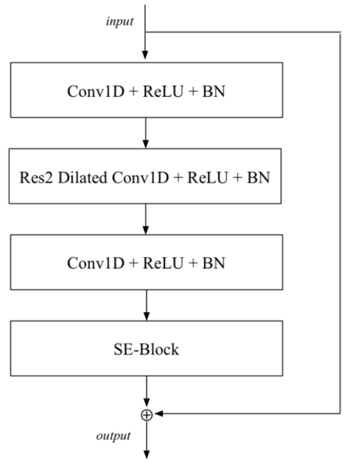

Fig. 1: The SE-Res2Block of the ECAPA-TDNN architecture

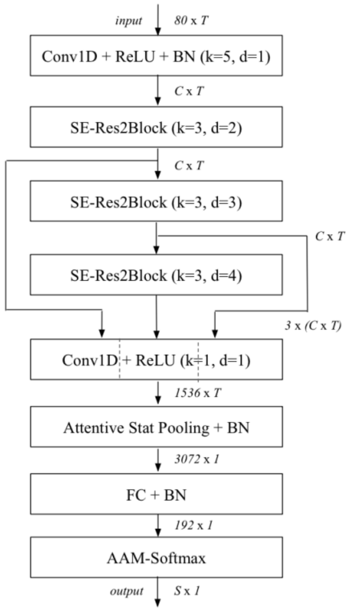

Fig. 2: Network topology of the ECAPA-TDNN

3. EXPERIMENTS

3.1. Database

The converted Speech dataset have been used for experimentation, The dataset has been created using two dataset one is Librispeech and other is VoxCeleb, former one is used as the source speaker and later one is target speaker.

3.2. Model

We have used ECAPA-TDNN as feature extractor. In the first layer we can choose if we want to work on raw audio or spectrogram, for the Experiments we have used spectrogram. The acoustic features are 80-dimensional log Mel-filterbank with sample rate 16k, window length 400 before going into deep network layers. The embedding size is 512, embedding size of 256 have also been used for the experimentation.

3.3. Training details

Similar to the baseline model we have used same network parametrs for training . Details are given below AdamW optimizer with cosine decay learning rate schedule. The initial learning rate starts from 1.0e-3, and the minimum learning rate is 1.0e-5. Same classifier is used like baseline model For the source speaker verification task which is ArcFace classifier. ECAPA-TDNN model have been trained on Voxceleb-2 [8] we have used this pre-trained ECAPA-TDNN and also trained ECAPA-TDNN from scratch on the subset of the training converted speech dataset.

4. RESULTS

We experimented the ECAPA-TDNN model using subset of training dataset. Firstly I trained the model in Train 1 dataset which gives the EER 60% when tested on the test set. As the model was only trained on the train 1 which was the smallest subset out of total 8 train dataset. After training of the model on Train 1, The checkpoint was saved and the checkpoint with lowest EER was used to again train the model with train1 dataset and subset of train2 dataset. which gives the EER to around to 57%. Again checkpoint was saved of the model having lowest EER.
Later ECAPA-TDNN was trained on the Train set of Train4, Train5, Train6 the checkpoint used was the previous checkpoint from train set of Train 1 and Train 2. We tested the model again on the testSet which gives the EER to 54%. Similarly we tested the model on different training set and calculated the EER the lowest EER we achieved was 50% on test set which was the lowest from all the other participant.

Results on Train Set		
Model	Dataset	EER %
ECAPA-TDNN	Train 1	69
ECAPA-TDNN	Train 1-2	57
ECAPA-TDNN	Train 4-5-6	54

5. CONCLUSION

In this paper, we have trained ECAPA-TDNN[1] on converted speech database of voxceleb and Librispeech dataset. Finding results on the dev and test set although the experimentation was not enough intensive to come close to the baseline model we trained the model on the subset of the converted dataset to find the performance of the model being exposed to reletively less data but Future work will aim to make the model performance better or to use other models for better performance.

.

6. REFERENCES

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," Oct. 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2650

[2] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, and M. Sahidullah, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," 09 2015.

[3] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof

2017 challenge: Assessing the limits of replay spoofing attack detection," 08 2017.

[4] A. Nautsch, X. Wang, N. Evans, T. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," 02 2021.

[5] W. Xingming, X. Qin, T. Zhu, C. Wang, S. Zhang, and M. Li, "The dku-cmri system for the asvspoof 2021 challenge: Vocoder based replay channel response estimation," pp. 16–21, 09 2021.

[6] M. Yousefi, N. Kanda, D. Wang, Z. Chen, X. Wang, and T. Yoshioka, "Speaker diarization for asr output with t-vectors: A sequence classification approach," pp. 3502–3506, 08 2023.

[7] Z. Li, Y. Lin, T. Yao, H. Suo, and M. Li, "The database and benchmark for source speaker verification against voice conversion," 2024.

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