FAST ADAPTATION OF PRETRAINED SPEAKER VERIFICATION SYSTEM FOR SOURCE SPEAKER TRACKING

Xiang Lyu, Yuxuan Wang, Tianyu Zhao, Huadai Liu

Alibaba Inc., Shanghai, China

ABSTRACT

Traditional speaker verification system aims at distinguish speaker identity in real world audio, and has achieved satisfying performance in many scenarios. However, it is also very vulnerable, and can be easily attacked by voice anonymization system. In this report, we describe how to fast adapt a pretrained speaker verification model to source speaker tracking task with pretrained feature and Lora[1] technique. It significantly reduce EER on voice anonymization system, as well as keep its performance in real world audio intact. Experiment on Attacker Challenge[2] shows that our system successfully reduce baseline EER by 32% in average, and achieve lowest EER in all voice anonymization system except T8-5.

Index Terms— Source speaker tracking, Pretrained speaker verification model, Pretrained feature, Lora

1. INTRODUCTION

Speaker verification system has gained substantial performance improvement due to the rapid development of neural network and large scale dataset. However, most research focuses on speaker verification system performance in real world audio, and its robustness against anonymization system remain untested.

Attacker Challenge is the succession of Voice Privacy challenge[1]. With the development of voice anonymization system, there has been concern about the abuse of these technique, and how to track the source speaker of voice anonymization speech when necessary. Based on the five SOTA and three baseline voice anonymization system from Voice Privacy challenge, Attacker Challenge focuses on reducing the EER in source speaker tracking task.

In this report, we design a speaker verification system which is initialized from pretrained speaker verification model. After finetuning with pretrained feature and Lora technique, we successfully reduced baseline EER by 32% in average, securing lowest EER on all voice anonymization system except T8-5.

2. PROPOSED SYSTEM DESCRIPTION

Attacker Challenge only provides the voice anonymization speech on LibriSpeech train-clean-360 subset, which has only 104,014 utterance and 921 speakers for each voice anonymization system. This amount of data is far from enough to train a speaker verification system from scratch. Thus, we employed a ResNet34 model[3][2] pretrained on VoxCeleb dataset for initialization.

[1]https://www.voiceprivacychallenge.org

[2]https://github.com/wenet-e2e/wespeaker/blob/master/docs/pretrained.md

2.1. Lora Adaptation

Real world audio can be changed significantly after processed by voice anonymization system, which makes them differ greatly in speaker embedding domain. Considering that real world audio has different data distribution with voice anonymization audio, we modify the ResNet34 model with Lora technique by adding a Lora module at each Conv2d layer. It should be noted that real world audio embedding will not be influenced by Lora module. To further increase the capacity of Lora module, we enhance it with re-param technique. The re-param Lora module design is shown in Figure 1.

2.2. Pretrained Feature Adaptation

Considering that limited amount of voice anonymization data may lead to poor generalization ability, we employ pretrained WavLM-large[4][3] feature for feature extraction. Due to the same concern in Lora adaptation design, we only extract WavLM feature for voice anonymization speech, and add it in the residual connection in each ResBlock. The modified ResBlock architecture is shown in Figure 2.

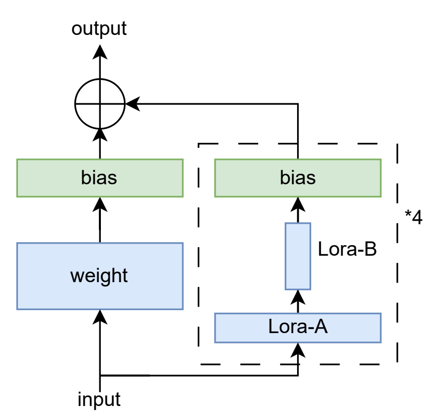

Fig. 1: Modified Conv2d module in ResBlock with Lora and re-param

3. EXPERIMENTAL RESULTS

3.1. Datasets

For real world data, we use VoxCeleb2 dev set and LibriSpeech train set for training. The LibriSpeech original train data contains 281,241 utterance from 2,338 speakers, and VoxCeleb2 dev set contains 1,092,009 utterance from 5,994 speakers.

[3]https://huggingface.co/microsoft/WavLM-large

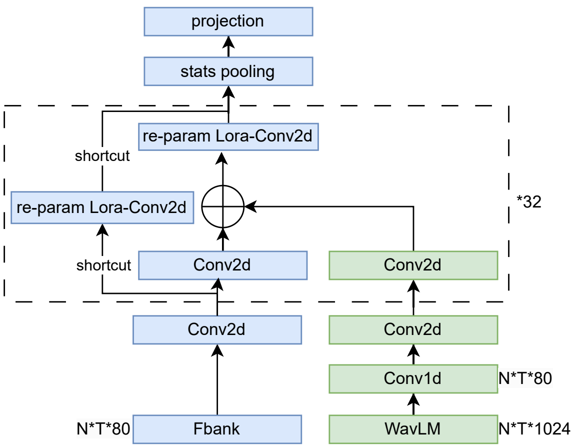

Fig. 2: Modified ResNet architecture with WavLM feature

For voice anonymization data, we only use the dataset provided by Attacker Challenge organizers. Each voice anonymization system contains 104,014 utterance from 921 speakers. We use WeSpeaker toolkit[4] for training.

3.2. Training Strategy

As the voice anonymization data is limited, we carefully design a three stage freeze-pretrain-finetune training strategy to prevent overfitting. WavLM model is fixed during training.

The pretrained ResNet34 projection layer has only 5,994 speakers from VoxCeleb2, while LibriSpeech dataset has an additional 2,338 speakers. In freeze stage, we initialize the modified ResNet34 model with pretrained ResNet34 backbone and projection weight. Then we freeze the backbone and only train the projection layer with real world data for 10 epochs.

In pretrain stage, we train the whole model with real world data and voice anonymization data for 70 epochs, thus equip it with ability to distinguish source speaker from voice anonymization speech.

Finally, we observe that in pretrain stage, loss and acc metric for each voice anonymization system fluctuate, which means that each voice anonymization system have different data distribution. In finetune stage, we train an additional 10 epochs with only voice anonymization data from each system and 1e-5 learning rate, yielding seven finetuned models for each system.

3.3. Ablation Study on B3

For convenience, we only conduct ablation study on B3 LibriSpeech test set.

model	EER		
	female	male	average
baseline	27.92	26.72	27.3
pretrained ResNet34	48	47	47.5
+freeze/pretrain	27.3	18.4	22.85
+finetune	25.5	16.9	21.2
+ASNorm	25.2	15.8	**20.5**

Table 1: Ablation study on B3 LibriSpeech test set

[4]https://github.com/wenet-e2e/wespeaker

In Table 1, original pretrained ResNet34 model achieve 47.5% average EER on LibriSpeech test trials, which means no capability to distinguish speakers in voice anonymization data. After pretrain stage, the average EER reduced to 22.8%, which is 17% relative improvement comparing with challenge baseline. After finetuning stage, the average EER is further reduced to 21.2%. We also perform ASNorm scoring with embedding from B3 voice anonymization training set as cohort. It further reduces average EER to 20.5%.

3.4. Evaluation Results

system	EER				
	dev		test		average
	female	male	female	male	
B3	27.9	19.5	25.2	15.8	**22.1**
B4	23.4	20.8	19.7	19.3	**20.8**
B5	28.6	26.7	26.9	24	**26.6**
T10-2	26	21.5	21.5	23.6	**23.2**
T12-5	29	28.5	26.6	24.4	**27.1**
T25-1	29.4	28.7	29.4	26	**28.4**
T8-5	31.5	26.7	27.5	25.1	**27.7**

Table 2: EER on LibriSpeech dev/test set for each system

For each system, we extract embedding using the final epoch from finetune stage, and calculate cosine similarity using ASNorm with top-n 300. It achieves 25.1% EER in average, which is 32% relative improvement comparing with challenge baseline. Submission result shows that our system achieves lowest EER in all voice anonymization systems except T8-5.

4. CONCLUSION

In this report, we describe how to train a source speaker tracking model from pretrained speaker embedding model, and enhance it with pretrained feature and Lora module. Evaluation results show that our training strategy successfully reduce EER on voice anonymization speech, securing lowest EER on all voice anonymization system except T8-5.

5. REFERENCES

[1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[2] Natalia Tomashenko, Xiaoxiao Miao, Emmanuel Vincent, and Junichi Yamagishi, "The first voiceprivacy attacker challenge evaluation plan," *arXiv preprint arXiv:2410.07428*, 2024.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.