Learning Device-invariant And Location-invariant Embedding For Speaker Verification Using Adversarial Multi-task Training

You-cai Qin, Qi-rong Mao, Zhong-chen Ma, Jing-jing Chen
School of Computer Science and Communication Engineering, Jiangsu University

优胜奖

Abstract

Deep speaker embedding model has achieved the satisfactory performance in close-talking speaker verification. However, in the real home environment, the device used to record voice may be different and the distance between speaker and device is constantly changing. These will make performance degradation. In this paper, a novel approach of adversarial multi-task training is proposed to solve the problem of device mismatch and location variation by learning device-invariant and location-invariant embeddings. A gradient reversal layer and a device-and-location classifier are added to the speaker validation model in order to build an auxiliary adversarial task. Experiments are conducted on the far-field text-dependent speaker verification database called HI-MIA, the proposed approach achieves 28% Equal Error Rate (EER) in close-talking enrollment task and achieves 9% EER in far-field enrollment task.

Proposed Method

1. Architecture

The proposed method for speaker verification (SV) consists of three main components, including a deep speaker embedding module, a speaker classification task module, and a device-and-location classification adversarial task module. During training, speaker classification task and device-and-location classification adversarial task jointly optimize the deep speaker embedding module. During testing, deep speaker embedding module transforms utterances into embeddings, and the similarity of embeddings between registered utterances and test utterances are computed to determine the identity of the speakers.

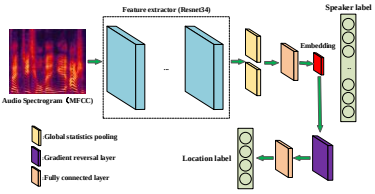

Fig. 1 Structure diagram of adversarial multitasking method

2. Deep speaker embedding module

Deep speaker embedding module includes three parts, a feature extractor, a pooling layer and a fully connected layer. The function of the feature extractor is to convert speech sequences into frame-level representations. ResNet can be able to solve the gradient disappearance problem by skip connections and has a good ability to learn representations. In this paper, ResNet34 is used as feature extractor. It is very important to translate frame-level representations into utterance-level representations by pooling function. GSP can provide more detailed information than global average pooling (GAP), so it is used as pooling function. Even though the frames of the speech signals are not as same in length, GSP is also able to translate them into the same dimension. The full connection layer outputs embeddings with discriminating information about the speaker.

3. Speaker Classification Task

In order for the deep speaker embedding module to learn speaker discriminative embeddings, the speaker classification task is significant as the primary task. Embeddings are mapped to the speaker categories and constrained by the speaker loss function. In this paper, two loss functions, SoftMax and Additive Angular Margin SoftMax (AAM-SoftMax) , will be investigated. AAM-SoftMax is an improvement on SoftMax with the addition of angular distance constraints. AAM-SoftMax reduces intra-class distances and increases inter-class distances. AAM-SoftMax is defined as follows:

$$L_{AAM-SoftMax} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s\left(cos\left(\theta_{y_i,i}+m\right)\right)}}{e^{s\left(cos\left(\theta_{y_i,i}+m\right)\right)}+\sum_{j=1,j\neq i}^{c} e^{s\left(cos\left(\theta_{j,i}\right)\right)}}$$

Where $\theta_{i,j}$ is the angle between the column vector W_j and x_i, s is a scaling factor and m is a hyperparameter controlling the margin.

4. Device-and-location classification adversarial task

To address the problem of device mismatch and location variation, the device-and-location classification adversarial task is added to the model as an auxiliary task. Using gradient reversal layer (GRL) to establish an adversarial process, deep speaker embedding module learns device-invariant and location-invariant embeddings.

It can be seen from Fig.1, device-and-location classification adversarial task module includes a GRL, a fully connected layer, and a device-and-location classifier. The input of gradient reversal layer are embeddings of the output of the deep speaker embedding module. In forward propagation, GRL is equivalent to a constant equality transformation. In back propagation, the gradients of device-and-location classifier is inversely propagated into the front network by GRL, thus enabling an adversarial loss similar to that of GANs. Device and location information in embeddings will be faded out during adversarial training. Therefore, deep speaker embedding module learns device-invariant and location-invariant embeddings.

5. Optimization

The loss of speaker classification task is minimized, while the loss of device-and-location classification adversarial task is maximized. Through the speaker classification task, the deep speaker embedding module can learn the discriminative representation of the speaker, and through the device-and-location classification adversarial task, the deep speaker embedding module can learns device-invariant and location-invariant representation.

The loss of the speaker classification task is $L_{AAM-SoftMax}$, denoted as $L_{speaker}$. The loss function of the device-and-location classification adversarial task is a combination of SoftMax and cross entropy, denoted as $L_{d\text{-}and\text{-}l}$, then the final loss function is defined as:

$$L_{final} = L_{speaker} + \alpha \times L_{d-and-l}$$

Experiment

1. Database

HI-MIA is a far-field text-dependent SV database. The database contains recordings of 340 people in rooms designed for the real home and far-field scenario. Recordings are captured by multiple microphone arrays located in different directions and distance to the speaker and a high-fidelity close-talking microphone. The circular microphone array contains 16 channels, each of which records in 16kHz, 16 bit. The close-talking microphone records high fidelity clean speech in 44.1kHz, 16 bit.

2. Front-end Processing And Experimental Setups

In this paper, Log Mel-filterbank energies (Fbank) features are used as an input to the model. The speech signals are framed in 25ms windows with a 10ms slide. Each audio is converted into 64-dimensional Fbank features. Voice activity detection (VAD) is used to remove silent segments in utterances to reduce the amount of data and improve speech quality. All utterances use energy-based VAD to remove silent segments.

The average length of the utterances in the database is 1 second, so the number of frames of speech features is randomly set from 100 to 140 in training. Because of insufficient data, the model is prone to overfitting. The validation set is used for training and the last training parameters are used for testing. The batch-size is set to 64. The initial learning rate is 1e-3, which drops to one-tenth of the original every 100 epochs.

3. Results

Table 1. Performances of the SV systems. 'TL' represents whether to use pre-trained models for transfer learning; 'AT' represents whether to use adversarial multi-task training.

Model	Task 1		Task 2	
	EER	minDCF	EER	minDCF
Resnet(Softmax,TL)	4.71%	-	3.70%	-
Resnet(Softmax)	6.26%	0.72	5.82%	0.57
Resnet(AAM-Softmax)	5.58%	0.56	5.01%	0.52
Resnet(AAM-Softmax,AT)	5.12%	0.54	5.08%	0.51
Resnet(Softmax,TL)	4.42%	0.43	4.01%	0.32
Resnet(AAM-Softmax,TL)	3.76%	0.35	3.46%	0.32
Resnet(AAM-Softmax,TL,AT)	**3.38%**	**0.35**	**3.36%**	**0.26**

It can be seen from the table that transfer learning is effective and can provide good initialization parameters. The performance of AAM-Softmax is better than that of Softmax, this proves that it is necessary to increase the distance between classes and reduce the distance within classes. In addition, our proposed multi-task confrontation training method can also improve performance.

Conclusion

In this paper, we propose a novel approach to adversarial multi-task training to address the device mismatch problem and location variation in speaker verification. A device-and-location classifier and a GRL are added to the model. The deep speaker embedding module can learns device-invariant and location-invariant embeddings by adversarial training. We also investigate two loss functions, SoftMax and AAM-SoftMax. AAM-SoftMax reduces intra-class distances and increases inter-class distances and gets better performance. Experiments are conducted on the HI-MIA database. The proposed method achieves a 28% relative improvement in EER over the baseline in the close-talking task and achieves a 9% relative improvement in EER over the baseline in the far-field enrollment task.

References

[1] Chiu, Chung-Cheng, et al. "State-of-the-art speech recognition with sequence-to-sequence models." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[2] Pandey, Laxmi , and R. M. Hegde . "Keyword Spotting in Continuous Speech Using Spectral and Prosodic Information Fusion." Circuits, Systems, and Signal Processing 38.6(2019)

[3] Rahulamathavan, Yogachandran, et al. "Privacy-preserving iVector-based speaker verification." IEEE/ACM Transactions on Audio, Speech, and Language Processing 27.3 (2018): 496-506.

[4] Chung, Joon Son, Arsha Nagrani, and Andrew Zisserman. "Voxceleb2: Deep speaker recognition." arXiv preprint arXiv:1806.05622 (2018).

[5] Novotny, Ondrej, et al. "On the use of dnn autoencoder for robust speaker recognition." arXiv preprint arXiv:1811.02938 (2018).

Contact

You-cai Qin
0086-18852853806
qiny1012@qq.com