A robust speaker recognition based on Data augmentation

Chenqi Li

Harbin Institute of Technology

Haoyuan Lu

Dalhousie University

Wei Wang (✉ wangwei_hitwh@126.com)

Harbin Institute of Technology <https://orcid.org/0000-0002-8441-1004>

Research

Keywords: speaker recognition, Data augmentation, spectral subtraction

Posted Date: March 9th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-16425/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

It is known that large-scale training data can get the better effect of recognition. However, it is difficult to collect a lot of labeled training data for speaker recognition. At the same time, the performance of speaker recognition is greatly influenced by environmental noise. In this paper, we use data augmentation by adding noise to get much training data and improve the robustness of speaker recognition. The experimental results demonstrate that data augmentation have the better performance improvement on Chinese-863 database.

1. Introduction

With the rapid development of computational power, Deep Neural Networks (DNN) like Convolutional Neural Networks (CNN) ^[1] and Recurrent Neural Networks (RNN) ^[2] have gradually developed in many artificial intelligence applications, including audio ^[3], text ^[4] and image ^[5]. Lei first applied deep neural networks to the field of speaker recognition and proposed a speaker recognition model based on i-vector and DNN, which has a relative improvement of 30% compared with the best performance of NIST SRE in 2012 ^[6]. After that, Kumar proposed DNN speaker recognition based on Embeddings, which has an improved performance of 45.5% compared with speaker recognition based on DNN ^[7]. Zhou stated that real-world problems are NP-hard, and if they are to be solved within polynomial time and using restricted number of samples, the answer can't be perfect, which means over-fitting can't be avoided^[8]. And, the neural networks have many parameters in model training and if the training data is too small, it is easy to cause over-fitting of the neural networks. For some tasks of speaker recognition, it is very difficult to collect substantial quantities of labeled data for training. So, we use data augmentation to increase the number and type of training data to prevent over-fitting.

Through the unremitting efforts of the researchers, it has been made significant progress in the research of speaker recognition. However, it is still a certain distance from research to practical applications because the performance of speaker recognition is greatly affected by environmental noise. Therefore, solving the robustness of speaker recognition becomes the critical step for its ability to move from research to application. However, in the real-world applications, the training data are usually expensive or even impossible to acquire. The reason why the robustness of speaker recognition is restricted is that the training database does not match the real data of application scene. In order to obtain the models that are most suitable with real-world applications, the most ideal situation is to use the data in the real world for training. However, it is not realistic to acquire large-scaled training data in real scenes. Data augmentation is used to get the training data that is close to the real scenes. The concept of data augmentation means augmenting the original data in order to generate new data.

There are many well-known effective speech augmentation methods, such as clipping ^[9], speed perturbation ^[10], noise addition ^[11], spec augment ^[12], etc. Based on the noise addition method, let (x, y) denote a training example as an observed input feature and label pair, data augmentation only augments the input feature x , yields augmented features $\tilde{x}$. While the augmentation is kept to be small

perturbations, the label remains unchanged. Therefore, it is an inexpensive method to acquire the amount of training data and improve robustness. So, we use data augmentation of the noise addition method to improve the robustness of speaker recognition.

2. The Flow Diagram Of Speaker Recognition

This paper adopts Mel Frequency Cepstral Coefficients (MFCC) feature extraction, where firstly, the preprocessed speech is framed into short frames, and Fast Fourier Transform (FFT) is calculated to obtain spectrum. Secondly, the Mel Triangle Filter bank is applied to the power spectra and the energy in each filter is summed up. Then, the logarithm of all filter bank energies is taken. Finally, Discrete Cosine Transform (DCT) is performed on the log filter bank energies to get MFCC. And we use Long Short-Term Memory Network (LSTM) model, which is a time-cycle neural network in order to solve the gradient disappearing problem of RNN. And the dropout layer is added after each layer of LSTM to improve the generalization ability of the network and avoid over-fitting problems. The average sequence of the output layer of LSTM is input as the depth into a Fully Connected Network (FC) for classification and finally the probability of each classification is calculated by the softmax function. The flow diagram of the system implementation is shown in Fig. 1.

3. Experimental Setup

3.1 Experimental Data Set

In order to evaluate the performance of data augmentation, experiments are carried out on Chinese-863 database. In the Chinese-863 database, the training set is composed of 60 male and 60 female speaker speeches and each of them is approximately 90s long. We use two methods to add the background noise in the speech data augmentation. One is the adding noise with the random signal-noise ratio (SNR) to the original training data, where SNR ranges from -10~-1dB, 0~10dB or 11~20dB. The other is the adding noise with the fixed SNR to the original training data, where SNR is fixed at 0dB, 10dB or 20dB. In addition to the original "clean" training list, the noise with random SNR is added to the training set, recorded as aug1 training set. And the noise with fixed SNR is added to the training set, recorded as aug2 training set. The aug1 and aug2 set are composed of about 360s speech per person. The test set has 720 true trails and 7200 false trails, where each endures 15s. In addition to the original "clean" testing list, noise with different SNR (0dB, 10dB, 20dB) and different noise types (factory2, factory1, babble) are added the test set. All speech files are wave format with sampling frequency at 8kHz and quantized with 16 bits. Speech streams are windowed into a sequence of short-term frame (25ms long) with 10ms overlaps.

3.2 Experimental Results and discussion

In order to construct the better speaker recognition, different experiments are carried out on the selection of feature parameters and model parameters.

Fixed on a 128-node LSTM network to train, the feature dimensions are selected 13 and 39 MFCC respectively. In Table 1, the experimental results clearly show that 39-dimensional MFCC is better than 13-dimensional MFCC. So, we use the 39-dimensional MFCC feature.

Table 1 Comparison of accuracy about the dimension of MFCC feature

MFCC feature dimension	Accuracy (%)
13-dimensional	93.89
39-dimensional	94.17

Fixed on 39-dimensional MFCC feature and 128-node LSTM network to train the model, the number of hidden layers of LSTM network is set one, two and three layers, respectively. In Table 2, the experiments clearly show that the best accuracy of speaker recognition exhibits at one hidden layer. Therefore, we use one-layer LSTM network.

Table 2 Comparison of accuracy about the number of different hidden layers.

the number of hidden layers	Accuracy (%)
one layer	94.17
two layers	91.67
three layers	90.69

Fixed on 39-dimensional MFCC feature and one-layer LSTM network to train the model, the number of nodes of LSTM network is tested using 128, 256 and 512 nodes respectively. In Table 3, the experiments clearly show that the best performance is got at the number of 512 hidden nodes. Therefore, we use 512-node network structure.

Table 3 Comparison of accuracy about the number of different hidden nodes.

the number of hidden nodes	Accuracy (%)
128	94.17
256	94.17
512	94.86

It can be seen from the above test results that we select hidden one hidden layer neural network with 512 nodes to train the model in the following experiment.

3.2.1 Robustness Test Based on Data augmentation

We use the traditional spectral subtraction ^[13] and data augmentation to compare the performance of speaker recognition. We use additive noise (factory2, factory1 and babble) to increase the amount of the existing training data. Among them, the spectral subtraction training model (clean-pj), the aug1 training model (aug1), and the aug2 training model (aug2) are used for evaluation. We respectively use “clean”, 0dB, 10dB and 20dB noise test set for testing.

In order to clearly show the experimental performance under 3 different SNR levels of factory2 noise environment. The different training models are evaluated under the factory2 noise (0dB, 10dB and 20dB SNR). The results are shown in Table4.

Table 4 evaluation of different training models under the factory2 noisy environment

SNR level	Training models		
	clean-pj	aug1	aug2
factory2_0dB	31.39	4.59	6.92
factory2_10dB	16.7	3.08	4.13
factory2_20dB	11.38	2.91	3.64

In table 4, the data illustrate that data augmentation (aug1 and aug2) are better than spectral subtraction (clean-pj).

Under the factory2 noisy environment (0dB) on three models, the results are shown in Figure 2, where the dotted lines (clean-pj(EER with 31.39%)), the dashed lines(aug1(EER with 4.59%)) and the dash-dot lines (aug2(EER with 6.32%)) describe the DETs.

Under the factory2 noisy environment (10dB) on three models, the results are shown in Figure 3, where the dotted lines (clean-pj(EER with 16.70%)), the dashed lines (aug1(EER with 3.08%)) and the dash-dot lines (aug2(EER with 4.13%)) describe the DETs.

Under the factory2 noisy environment (20dB) on three models, the results are shown in Figure 4, where the dotted lines (clean-pj(EER with 11.38%)), the dashed lines (aug1(EER with 2.91%)) and the dash-dot lines (aug2(EER with 3.64%)) describe the DETs.

According to the above DETs, it is obvious that data augmentation can improve the robustness of speaker recognition under factory2 noise, and aug1 model has the better performance than aug2 model.

In order to clearly show the experimental performance under 3 different SNR levels of factory1 noise environment. The different training models are evaluated under the factory1 noise (0dB, 10dB and 20dB SNR). The results are shown in Table 5.

Table 5 Evaluation of different training models under the factory1 noisy environment

SNR level	Training models		
	clean-pj	aug1	aug2
factory1_0dB	45.67	26.65	38.19
factory1_10dB	27.97	9.29	13.65
factory1_20dB	17.75	3.73	5.39

In table 5, the data illustrate that data augmentation (aug1 and aug2) are better than spectral subtraction (clean-pj).

Under the factory1 noisy environment (0dB) on three models, the results are shown in Figure 5, where the dotted lines (clean-pj(EER with 45.67%)), the dashed lines (aug1(EER with 26.65%)) and the dash-dot lines (aug2(EER with 38.19%)) describe the DETs.

Under the factory1 noisy environment (10dB) on three models, the results are shown in Figure 6, where the dotted lines (clean-pj(EER with 27.97%)), the dashed lines (aug1(EER with 9.29%)) and the dash-dot lines (aug2(EER with 13.65%)) describe the DETs.

Under the factory1 noisy environment (20dB) on three models, the results are shown in Figure 7, where the dotted lines (clean-pj(EER with 17.75%)), the dashed lines (aug1(EER with 3.73%)) and the dash-dot lines (aug2(EER with 5.39%)) describe the DETs.

According to the above DETs, it is obvious that data augmentation can improve the robustness of speaker recognition under factory1 noise, and aug1 model has the better performance than aug2 model.

In order to clearly show the experimental performance under 3 different SNR levels of babble noise environment. The different training models are evaluated under the babble noise (0dB, 10dB and 20dB SNR). The results are shown in Table 6.

Table 6 Evaluation of different training models under the babble noisy environment

SNR level	Training models		
	clean-pj	aug1	aug2
babble_0dB	33.87	19.06	20.41
babble_10dB	20.69	7.24	8.59
babble_20dB	13.86	3.88	4.69

In table 6, the data illustrate that data augmentation (aug1 and aug2) are better than spectral subtraction (clean-pj).

Under the babble noisy environment (0dB) on three models, the results are shown in Figure 8, where the dotted lines (clean-pj(EER with 33.87%)), the dashed lines (aug1(EER with 19.06%)) and the dash-dot lines (aug2(EER with 20.41%)) describe the DETs.

Under the babble noisy environment (10dB) on three models, the results are shown in Figure 9, where the dotted lines (clean-pj(EER with 20.67%)), the dashed lines (aug1(EER with 7.24%)) and the dash-dot lines (aug2(EER with 8.59%)) describe the DETs.

Under the babble noisy environment (20dB) on three models, the results are shown in Figure 10, where the dotted lines (clean-pj(EER with 13.86%)), the dashed lines (aug1(EER with 3.88%)) and the dash-dot lines (aug2(EER with 4.69%)) describe the DETs.

According to the above DETs, it is obvious that data augmentation can improve the robustness of speaker recognition under babble noise. And aug1 model has the better performance than aug2 model, because aug1 model includes more different SNR information than aug2 model.

It can be seen from all test results that, under different noisy environment, data augmentation is always better than spectral subtraction in the robustness of speaker recognition. And the experiments show that data augmentation is an effective method for having the better performance improvement. However, there are still some differences between the real scenes data and data augmentation data.

4. Conclusion

This paper use data augmentation by adding noise to collect the large-scale labeled training data and improve the robustness of speaker recognition. Experiments are tested respectively on spectral subtraction and data augmentation under factory2 noise, factory1 noise, and babble noise environment on Chinese-863 database. The test results show that data augmentation of adding noise can improve the robustness of speaker recognition, where the random adding background noise is better than the fixed adding background noise.

Abbreviations

DNN: Deep Neural Networks; CNN: Convolutional Neural Networks; RNN: Recurrent Neural Networks; FFT: Fast Fourier Transform; MFCC: Mel Frequency Cepstral Coefficients; DCT: Discrete Cosine Transform; LSTM: Long Short-Term Memory Network; FC: Fully Connected Network; SNR: signal-noise ratio.

Declarations

Acknowledgements

Not applicable.

Authors' contributions

All authors have contributed to this work. All authors read and approved the final manuscript.

Funding

This research is supported by the National key research and development plan (No.2017YFB0803001), National Science Foundation of China (No. 61571144) and the Fundamental Research Funds for the Central Universities (No. HIT.NSRIF.2019082).

Availability of data and materials

Please contact author for data requests.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Harbin Institute of Technology, Harbin, China. ²Dalhousie University, Halifax, Canada.

References

1. LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." Neural computation, Vol .1.4, 1989, pp.541-551.
2. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation, Vol. 9.8, 1997, pp.1735-1780.
3. Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." International conference on machine learning, 2016, pp. 173-182.
4. Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078,2014.
5. Krizhevsky, Alex, IlyaSutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, 2012, pp.

1097-1105.

6. Lei, Yun, et al. "A novel scheme for speaker recognition using a phonetically-aware deep neural network." In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, 1695-1699.
7. Nandwana, Mahesh Kumar, et al. "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings." In Interspeech, 2018, 1106-1110.
8. Zhihua Zhou. Machine Learning. Tsinghua university press, 2016.
9. Panayotov, Vassil, et al. "Libri speech: an ASR corpus based on public domain audio books." In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, 5206-5210.
10. Ko, Tom, et al. "Audio augmentation for speech recognition." In Sixteenth Annual Conference of the International Speech Communication Association, 2015.
11. Ko, Tom, et al. "A study on data augmentation of reverberant speech for robust speech recognition." In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, 5220-5224.
12. Park, Daniel S., et al. "Spec augment: A simple data augmentation method for automatic speech recognition." arXiv preprint arXiv:1904.08779, 2019.
13. Joshi V, Prasad N V, Umesh S. "Modified Mean and Variance Normalization: Transforming to Utterance-Specific Estimates." Circuits, Systems, and Signal Processing, 2016, 35(5):1593-1609.

Figures

原始wav文件

MFCC

提取39维MFCC特征

LSTM层

Dropout

Dropout层

FC

全连接层

Softmax

输出层

识别结果

Figure 2

The flow diagram of system implementation

Figure 3

DETs of three models under the factory2 noisy environment (0dB)

Figure 6

DETs of three models under the factory2 noisy environment (10dB)

Figure 8

DETs of three models under the factory2 noisy environment (20dB)

Figure 10

DETs of three models under the factory1 noisy environment (0dB)

Figure 12

DETs of three models under the factory1 noisy environment (10dB)

Figure 14

DETs of three models under the factory1 noisy environment (20dB)

Figure 16

DETs of three models under the babble noisy environment (0dB)

Figure 17

DETs of three models under the babble noisy environment (10dB)

Figure 19

DETs of three models under the babble noisy environment (20dB)