Content-Independent Speaker Recognition System based on Neural Networks

Hangyu SONG† Hiroshi WATANABE†
†Graduate School of Fundamental Science and Engineering, Waseda University

Abstract: In this research, we tried to build the content-independent speaker recognition system with neural networks. We built the speaker recognition system with RNN, CNN and RNN-CNN to distinguish the voice of 2 speakers. The results showed that for all of the 3 networks, the accuracy is obviously higher than random choice. It is proved that neural network is an effective approach to extract the features of voice.

1. Introduction

In the past, many methods have been developed for speaker recognition. Voiceprint recognition has been widely applied in the speaker recognition fields. However, for most of the methods, the input voice need to be certain content. In recent years, neural networks in the machine learning field developed rapidly. Hinton's group used neural networks for speech recognition [1]. This suggests that it is possible for neural network to extract features from voice.

In the daily life, people can distinguish the speakers by their timber. This implies that timbre must be a recognizable feature of voice. By building different neural networks, we verified that the neural network can be used to recognize the speaker from the voice.

2. Method

We aim to distinguish 2 speakers, as the first step of the research, 3600 voice samples with speaker information labels of 2 speakers (1800 samples for each speaker) are prepared. 2800 samples are used as used in the training of networks, while rest 800 samples are used in the evaluation.

To extract the frequency domain information, MFCC (Mel-Frequency Cepstral Coefficient) was a popular method in the audio processing field [1]. In this research, for each voice sample, the voice was sampled 8 times, with window length 0.5s and step length 0.25s. For each fragment of the sample, the MFCC is computed with window length 20 ms and step length 10 ms. Considering the highest frequency for human voice is at about 1.1 kHz, the frequency range of MFCC is set from 0 to 2.2 kHz. Thus, an MFCC graph is made for the sample. The graph is used as the input of the network.

Considering that voice was a time-related value, the RNN is used as one choice of the network [2]. At the same time,

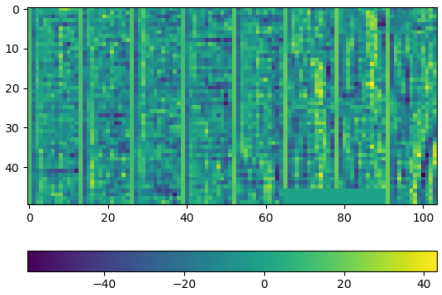

Figure 1 A Sample of MFCC Graph

in fact, the graph contains both frequency domain information and time information (location on graph), the graph can be also treated as an image. CNN was a popular method in image processing, so, CNN is also used as a choice of the network. The combination of RNN and CNN is the third choice of the network.

3. Experiment Results

3 networks were built and trained with TensorFlow. As a result, all of accuracy of networks was obviously higher than 50%.

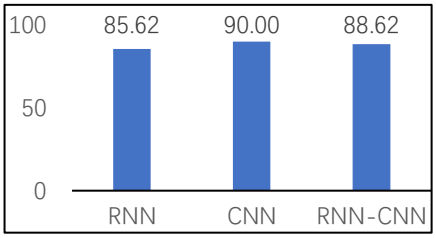

Figure 2 Accuracy of Networks (in percentage)

4. Conclusion

In this paper, we propose the content-independent speaker recognition system with neural networks, specifically RNN, CNN, and RNN+CNN.

Reference

[1] A. Graves, A. Mohamed and G. Hinton, Speech Recognition with Deep Recurrent Neural Networks, *CoRR*, abs/1303.5778, 2013

[2] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur and A. Mertins, Audio Scene Classification with Deep Recurrent Neural Networks, *CoRR*, abs/ 1703.04770, 2017