Speaker Recognition System based on Deep Learning

Hangyu SONG † Hiroshi WATANABE †

†Graduate School of Fundamental Science and Engineering, Waseda University

Abstract: In this paper, we introduced a speaker recognition system based on the deep learning. The system can be used to recognize 2 speakers regardless the content. By doing the one-dimension convolution on audio waveform, the system extracts feature from wav file. The accuracy reached about 60.6% with recall at 78.3%. The result shows that the system still needs to be optimized.

1. Introduction

With the development of smart equipment, the recognition system also needs to be built to meet the requirement of equipment. Nowadays, artificial intelligent has been used in many recognition systems, especially for the neural network technology. For example, CNN (convolutional neural network) technology was widely used in face recognition for smart phone. The speaker recognition system will also be an important part in the future smart systems.

It was a difficult task to deploy neural network in the past. However, the development of hardware and programming tools make it possible to deploy pre-trained neural networks on various of equipment. In this paper, we tried to make a speaker recognition system based on neural network.

2. Past Works

In the past works, voiceprint recognition technology was used in the recognition of speaker. However, it was a content-related system, i.e. speaker need to say certain content to pass the test.

In a research in 2013, neural network was used for speech recognition. [1] This suggests that, neural network can be used in recognition of vocal sounds. So, we made a series research in 2018 to verify if it is possible to recognize speakers with MFCC features. [2] The results illustrated that, CNN and RNN (recurrent neural network) both worked well in classification of 2-speaker-recognition task.

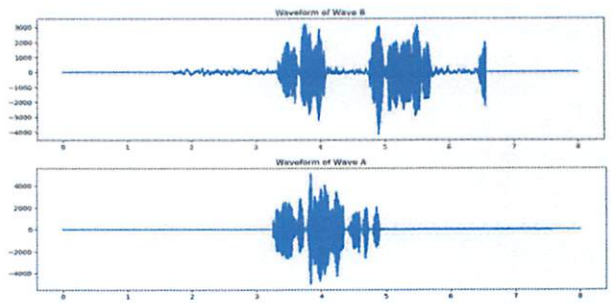
In 2016, DeepMind released a voice generation model called WaveNet, which generate human voice based on one dimension convolution. [3] This suggests that, 1-D convolution maybe a good choice to extract features from waveform of human voice.

3. Purpose

As a following up research of the research in 2018, we require the system in this research to recognize the mix of voice of 2 speakers. The system should give out a timed label for a waveform file.

4. Method

In this research, we used 4153 wave samples of 2 speakers



[0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0]
[0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0]

Figure 1 One Sample and Ground Truth of Waveform

(about 2000 samples for each speaker) to make training and evaluating samples. 3153 voice samples (about 1500 for each speaker) were used to make 5000 training sample by random mixing. The rest voice samples were used to made 1000 evaluating samples. The length of every training/evaluating sample is 8 seconds. Every 0.5 seconds, the network gives out a sigmoid label. The threshold for label is 0.7, which means for every window of waveform, if the possibility of speaker A/B is speaking is larger than 70%, it will be labeled with positive, otherwise, is negative.

The network was built with 1-D convolution layers and fully connected layers. The residual block was also introduced in network. In order to sketch the sight view of convolutional kernel, dilated convolution was used in convolutional parts.

5. Results

After the training of 500 epochs, the loss of network remains stable. Evaluation result got an accuracy at 60.6%. The recall of network was 78.3%. it turned out that, for the mixing waveform part, the number of false positive results was considerably large. This suggests that, for 1-D convolution, it may be difficult to distinguish the frequency features of sound.

6. Future Work

In the future work, we will try to use short-time Fourier transform spectrogram to extract frequency features of voice. We will also try to recognize voice of more speakers.

Reference

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

[2] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, WaveNet: A Generative Model for Raw Audio, *CoRR*, abs/1609.03499, 2016

[3] H. Song and H. Watanabe, Content-Independent Speaker Recognition System based on Neural Networks, *IEICE General Conference*, D-12-9, Mar. 2019