Journal of Electronic Research and Application

Research Article

A Speaker Identity Recognition System based on Deep Learning

Yili Shen*

Shanghai Weiyu High School, Shanghai, China

Abstract: This paper describes a branch of pattern recognition and lies in the field of digital signal processing. It is a speech recognition system of identifying different people speaking based on deep learning. In brief, this method can be used as intelligent voice control like Siri.

Keywords: Speech recognition, Intelligent, Signal processing

Publication date: September, 2019

Publication online: 30 September, 2019

****Corresponding author:*** Yili Shen, liaoquanneng@ivygate.cn

1 Introduction

With mankind entering the information age, information recognition technology has become a key research area in the scientific community. It is far from enough to meet people's needs to recognize targets through fingerprints and facial images. Therefore, speech recognition has become the focus of attention. Speech recognition technology has been proposed long before the invention of computer. With the rise of AI in the 21st century, speech recognition technology is gradually maturing. Speech recognition technology can be applied in many fields, such as automatic data entry, machine translation, security system, Siri and other human-computer interaction systems.

In traditional voiceprint recognition, we must extract all speech segments, and then compensate the speech information based on various intelligent algorithms. The traditional algorithm performs well in anti-noise, especially in the case of low signal-to-noise ratio. Moreover, this method is very complex and inefficient, especially when the speaker is in a complex environment. In addition, the accuracy often falls short of our requirements.

In this project, TensorFlow, which is widely used in machine learning applications, is used as a Deep Learning framework to implement the model. In the whole training process, the program will optimize the weights and deviation sets of each layer of the neural network to minimize the loss function, so as to achieve the purpose of accurate voiceprint speech recognition.

2 Experimental verification

2.1 Data preparation

In order to get a high accuracy rate, we collected 200 audio files from each of our members. Each file contains 10 seconds of audio so that we can make sure the file is usable after we use specific functions to edit it.

2.2 Data processing

To make sure the audio file is able to read in programs and functions, we use a code 'audio[:,0]' to convert the file from a two-dimensional matrix to a one-dimensional matrix. Then we read the file into Voice Processing, a program that cites characteristics of frame-shift of an audio file, to convert the audio from a one-dimensional matrix to a frame matrix. Frame-shift is an interval between two frames. In an audio file, the last part of one frame is the front part of the next frame, the overlapping part is called frame-shift. So if we want to separate frames from a file, we need to delete frame-shift. A frame matrix shows the details and features of an audio file.

2.3 Valid section extracting

Each of frame matrix will be cited in the third program: Valid Detect. Furthermore, the program Voice processing also builds two functions: Zero-crossing rate and Short-time energy. These functions are used to detect silence clips of audio. Program Valid Detect uses Double-threshold testing to extract a valid part of a file. In this project, it is necessary to build a comparable database to produce a high accuracy, so we have to make sure the section of the audio is a valid section rather than a fragment of noise. This is how Double-threshold testing works. The first judgment uses short-term energy to make a rough judgment according to the outline of short-term energy. The threshold T1 is chosen to intercept the starting point A and the ending point B, which is outside the AB section.

Then the threshold T2 is determined by using average energy. Searching from point A to left and point B to right, we can find the CD points intersecting with threshold T2. The CD segment is the voice segment determined by double threshold method.

The second judgment is based on the short-term energy zero-crossing rate. Search from C point to left and D point to right, and find EF point whose zero-crossing rate is lower than threshold T3.

EF segment is the beginning and ending segment of audio. Threshold T3 is determined by zero-crossing rate of background noise. We need to use the start point and endpoint, produced by Double-threshold testing, to edit the audio to a valid section.

2.4 Features extracting

Since everyone is different, everyone equips different characteristics of voice, we have to extract these characteristics in order to separate people from people. MFCC (Mel-scale Frequency Cepstral Coefficients) is a method. The valid section we get from the program Valid Detect would be utilized in the fourth program Mfcc. Moreover, we have to cut the valid audio file into 2.5 seconds so the computer could run it quicker. Through this function, we get a file which cut into 2.5 seconds and a matrix containing the characteristics of the file, the matrix is needed to be a sample to train the project.

2.5 Reshaping

After we read 1200 audio files (6 members, 200 files per member), we get a three-dimensional matrix, the shape is [200, 1024, 16] (originally the second element is a number bigger than 1024 instead of 1024, but for a binary computer, 1024(32*32) is the best data it could deal with). The matrix [1024, 16] means, for example, Member 1 has 200 files and he named his first audio file as ‘1_1.wav’. Now he uses this file to run the program and he gets his matrix [1024, 16], produced by MFCC, which means there are 16 ‘pictures’, each of ‘picture’ belongs label ‘1’ and contains 1024 data. However, a ‘picture’ equips length and width, so we have to reshape 1024 to two numbers. Eventually, we classified the files in a ratio of 4:1 so we have a matrix contains 960 files in file folder ‘Train’ and a matrix contains 240 files in file folder ‘Test’. These matrixes would be used in machine learning stage.

2.6 Matrix processing

However, for the function, it could only deal with a matrix with a formal ‘picture’ liked structure, so we have to reshape the matrix to another type of matrix. The structure of ‘picture’ is [(The number of pictures), (The width of a picture), (The length of a picture), (The number of channels)]. However, the structure of the original matrix is [(The width of a picture), (The width of a picture), (The number of channels), (The number of pictures)]. So, elements in the new matrix should be [3,0,1,2]. What we need to do is reshape the original matrix to the right matrix so that the program could run it.

2.7 Convolutional network

Convolution, as an important function to extract characteristics of matrix, is usually be utilized in machine learning project. Generally speaking, Convolution collects and amplifies the characteristics from a matrix. Through this function, program could separate data from data and increase the rate of accuracy.

3 Conclusion

Due to our originality, there is our innovation that need to be protected in our method. The network architecture features our research. It is not complicated, nevertheless, it has incomparable ability to suit for all kinds of environments with sufficient data provided. If the extracted features are enough to support its learning, our invention can distinguish the identity of the speaker with eminent accuracy.