EL-GY 9123-I Intro to Machine Learning

Project Report

Speech Recognition

Project:	Speech Recognition
Student:	Shuai Zhang
	Yue Su
ID:	sz1950 N15408473
	ys3231 N11043598
Instructor:	Prof.Sundeep Rangan

2017. 12

1. Background

Language is one of the most important communication tools for human beings. It's also one of the most important human information sources. Let the machine has the ability of understanding human oral language is a very interesting topic. On one hand, it will make machine more intelligent. On the another hand, it will make the machine have one more approach to help human beings. In 1952, AT&T Bell Labs first develop the system Audry which has the ability to identify ten English numbers. Since that, the speech recognition technology has been developed rapidly. This year, Google, Apple and Microsoft have all developed their audio assistant basing on speech recognition. This technology not only helps drivers to make a phone call safely, but also gives blind people a way to operate machine, to make them have the opportunity to feel the convenience brought by modern technology.

2. Goal

The goal of this project is to build up a real time system which can recognize English numbers and some simple commands to enable us to operate the computer and this speech recognition system by voice.

The implementation of this project is Matlab. That is because speech recognition needs to deal with audio signal and Matlab is a very powerful tool to process digital signal with its DSP ToolBox. Also, it is efficient to calculate matrix.

3. Procedure

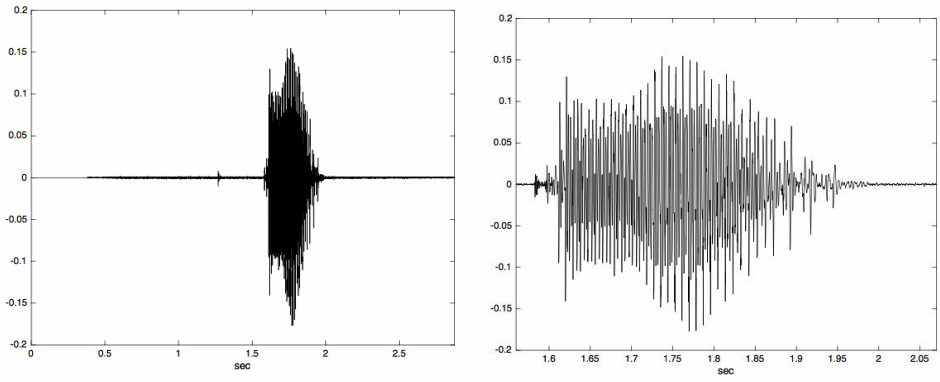
We implement this project in three steps. This three steps is commonly seen in machine learning implementation.

a) Preprocessing

The purpose of preprocessing is to minimize the influence of irrelevant element and provide a signal easy to extract feature. In other words, no matter signals are gathered in what circumstance, the format of all after going through the preprocessing step should be as similar as possible.

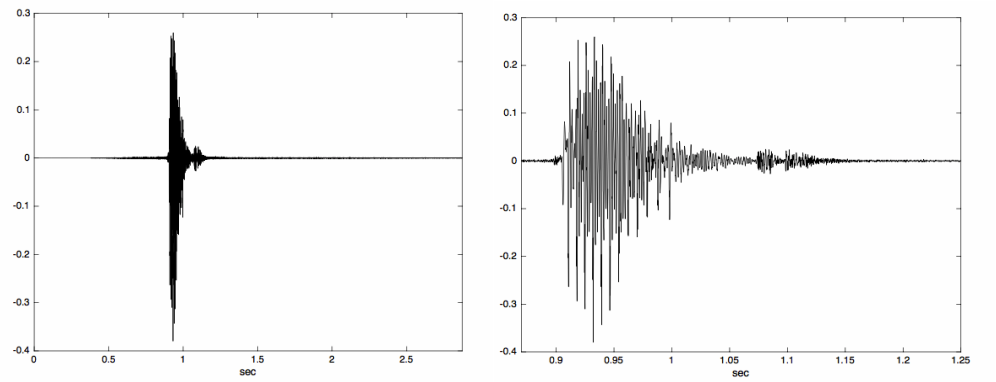
First, We let the raw signal go though a band pass filter which has passing band from 80 Hz to 3400 Hz, the human voice frequency limitation. Noisy produced by facilities in buildings such as air conditioner are usually under 80 Hz.

Then, We extract valid signal from the raw. In other words, find the start point and the end of a speech. To do this, We first separate the signal into frames. Those frames are of length 240 samples with overlap of 1/4. Then, We calculated its short time energy and set the start point threshold and the ending point threshold. After that, We compare the short time energy of each frame with the two thresholds. What is needed to pay attention is that the beginning threshold needs to be higher than the ending threshold. Because most words contain more than one syllable. We need to make sure the fragment contain all syllables of this word. Also, we need to set a silence frame number. Only if number of silence frames after a speech larger than the set number we can consider it is the end of this speech. By doing this, we can get a fragment of a phrase but not only a word. This enable the system to recognize some phrases commands like ‘open calendar’.



(a) (b)

Figure 1 (a)signal ‘three’ before extraction(b) signal ‘three’ after extraction



(a) (b)

Figure 2 (a)signal 'six' before extraction(b) signal 'six' after extraction

b) Feature extraction

Generally speaking, features extraction gets as much as identification from the raw data and in the same time remove interference. In the step of feature extraction, we use the Mel-frequency cepstral coefficients(MFCC). MFCC is a way to get identification based on the human ears characteristic and transfer every wave frame to a vector that represent its features. It has been showed that MFCC has higher recognize correctness than other methods. So it's more suitable to implement human voice recognition.

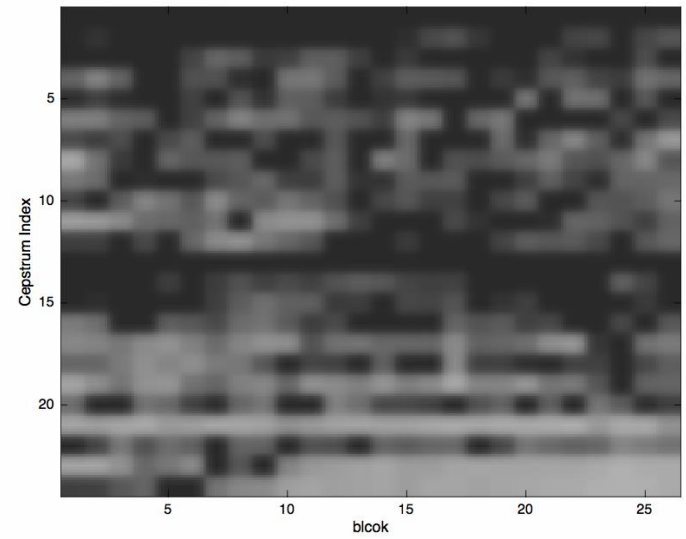


Figure 3 MFCC of signal 'six'

Typically, the are 6 steps to implement the MFCC. They are preprocessing, frame the signal, compute the power spectrum, apply Mel filter bank, take the logarithm

and compute the discrete cosine transform.

After preprocessing, we frame the signal into short frame. The purpose of this step is to avoid information losing which results from the continuous changing of speech signal.

Then, in order to minimize the Gibbs effects, we add Hamming windows to every frame. Then we take the FFT of each frame. After the FFT, the linear spectrum calibration of each frame is turned into Mel calibration which is conformed to the human ear hearing characteristics. Then it's filtered with a group of triangle band pass filter. Here, the number of triangle band pass filter is 24. The relationship of Mel calibration and linear spectrum calibration is as follow,

$$Mel(f) = 2595 * log_{10}(1 + \frac{f}{700})$$

In the end, we make a discrete cosine transformation of the 24 results of get by the triangle band pass filters. The discrete cosine transformation formula is as follow[1],

$$C_m = \sum_{k=1}^{N} E_k \cos\left(m(k - \frac{1}{2})\frac{\pi}{N}\right), m = 1, 2...L$$

c) Feature matching

In the feature matching step, we use dynamic time warping (DTW) algorithm, which effectively solves the problem of speech signal feature extraction and unequal-length speech matching. In addition, there are also many other feature matching methods like hidden Markov model (HMM) and HTK, but these statistics-based recognition technology requires a large number of voice materials to extract statistical features. Relatively speaking, the DTW is much easier to implement, and it can be used in this project with sufficient accuracy in the recognition of specific words even phrases with improvement of preprocessing step.

DTW is a non-linear regularization of time regularization combined with distance measurements. The idea of the algorithm is to evenly elongate or shorten the unknowns until the match the length of the reference pattern. In this process, the

time axis of the unknown speech will be uniformly distorted or bent to align its features with the model features.

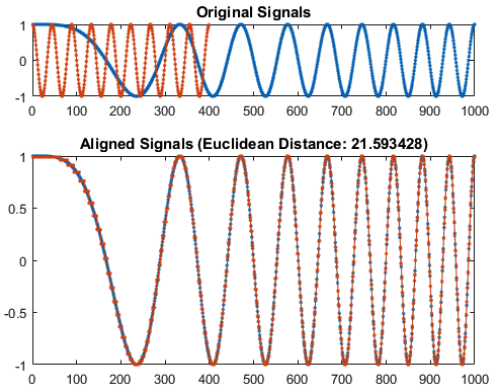


Figure 4 an example for DTW[2]

4. Structure

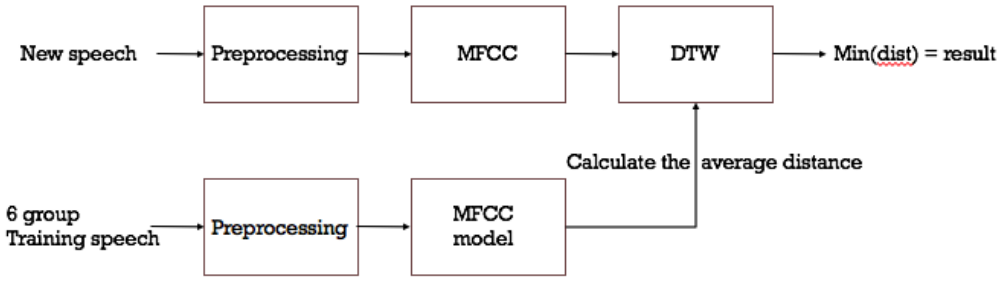


Figure 5 structure of this system

Figure 5 shows the structure of this speech recognition system. There are two path. The upper path is the test path. The lower path is the model path. The two path all experience the same preprocessing and feature extraction. After step two, the 6 training models were saved as a .mat file as a reference model. The purpose of using 6 models is to minimize the influence cause by variant accent due to dtw's small database. In the main function the raw test speech is record and send to the speech recognition function sr_real_time_fun(). The reference model is sent to it as well. In the speech recognition function, the function find_fragment() and mfcc() are called to preprocessing the raw test signal. After that, function dtw() is called to compute the distance between the test speech and 6 reference models. The one with the shortest average distance will be the result as a return of sr_real_time().

5. How to use

For recognition:

- Open file fun_main.m and press run to use this system.
- Speak zero to nine to recognize number.
- Speak ‘open calendar’ to open calendar.
- Speak ‘open calculator’ to open calculator.
- Speak ‘over’ to end this program.

For training:

- Save training speech to director data.
- Open train.m and modify file path.
- Run it and save the reference model as .mat
- Use main file to load .mat file and to recognize.

Notice:

- The system will run for 180 seconds by default unless receives command ‘over’
- Commands are only worked for Mac OSX.
- For windows, please modify function system().
- ‘over’ pointed to number 10.
- ‘open calendar’ pointed to number 11.
- ‘open calculator’ pointed to number 12.

6. Application and limitation

For the number recognition part, this system can be modified to a automatic telephone answering service system. For example, Amazon and AT&T’s automatic custom service. It can also be modified to a audio dial system for driver so that they do not need to look at their cellphone while driving.

For the commands recognition part, it can be modified to a audio assistant to help blind people to operate computer or to make everyone lives a more convenient life.

Of course, this system still has many limitations. First, it is implemented by Matlab. It’s not a free software so it’s portability is not as good as other language. It will be better if it’s rewrite by python or C++. Second, the feature matching algorithm determined its bottleneck: the more item to recognize, the less correctness. Because as the number of recognition item grown, the distance

between each other will get closer and closer. The way to solve this is to change feature matching algorithm such as HMM.

7. Github

https://github.com/YueSugithub/Speech-Recognition

Reference

[1] Huapeng Wang, Hongchen Yang. Research on Voiceprint MFCC Features' Extraction [J]. Journal of the People's Public Security University (Natural Science Edition), 2008, 55(1): 28~30.

[2] Signal Processing ToolBox Documentation https://cn.mathworks.com/help/signal/ref/dtw.html

[3] Yaqin Liu, Aijuan Zhang, The Study of Several Speech Recognition Feature Parameters. COMPUTER TECHNOLOGY AND DEVELOPMENT, Vol. 19 No. 12Dec.2009..

[4] Xuedong Huang, Li Deng Handbook of Natural Language Processing. Microsoft, Page 339, 2009-9-9.