Project Report of Group12

Isolated Word Recognition System

Author: Jiahao Hu. Zhuo Liu.

Division of work:

Jiahao Hu: Algorithm, implementation, report, PowerPoint.

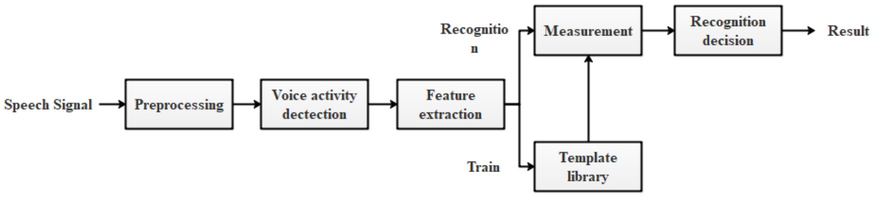
Zhuo Liu: MATLAB GUI.

Introduction

The goal of speech recognition is to make the machine "understand" the natural language of human beings, that is, to recognize the content of speech accurately in various situations, and then to execute the various intentions of human beings according to their information. Speech recognition has broad application prospects in industry, military, transportation, medicine and civil areas, such as voice-controlled telephone exchange, voice dialing system, various voice and voice services (stock information, weather forecast, etc.), intelligent toys, voice call center, etc. Speech recognition technology will greatly improve the human-machine interface, improve the automation of information processing, and have a huge society. Economic benefits. Speech recognition is rapidly developing into one of the key technologies of "changing future human life style".

Structure of Speech Recognition System

The typical implementation scheme of speech recognition system is shown in the figure. Input analog initial message and so on need not pass through the dam surface sentences, including pre-filtering, sampling and quantization, windowing, endpoint detection, pre-emphasis and so on. After speech signal is preprocessed, the next important step is feature parameter extraction. The requirements for characteristic parameters are as follows: (1) The extracted feature parameters can effectively represent speech features and have good discrimination. (2) There is good independence between the parameters of each order. (3) In order to ensure the real-time realization of speech recognition, the feature parameters should be calculated conveniently and efficiently.



In the training stage, the feature parameters are processed to obtain a model for each entry, which is saved as a template library. In the recognition stage, the voice signal passes through the same channel to get the voice parameters, generates a test template, matches with the reference template, and takes the reference template with the highest matching score as the recognition result. At the same time, it can improve the accuracy of recognition with the help of a lot of prior knowledge.

Speech Recognition Feature Parameters

Speech feature extraction is the basis of speech recognition and a key technology related to the performance of recognition system. The selection of speech feature vectors will directly affect the performance of recognition system. The basic idea of feature parameter extraction is that the preprocessed speech signal is transformed once to remove the redundant part, and the feature parameters representing the essence of speech are extracted. At present, the feature vectors commonly used in speech recognition can be divided into the following two types: (1) Cepstrum coefficients based on LPC: (2) Spectral cosine transform analysis based on FFT. The first one is LPCC, which is a linear predictive cepstrum coefficient based on Durbin algorithm. The second type has Mel frequency cepstrum coefficients based on Mel scale. The analysis of Mel frequency cepstrum parameters focuses on the auditory mechanism of the human ear and analyses the frequency of speech based on the results of auditory experiments. Compared with cepstrum analysis based on linear prediction, its prominent advantage is that it does not depend on the assumption of full-pole speech to produce splitting, and shows stronger robustness in noise environment. A large number of studies show that MFCC is superior to LPCC in terms of recognition effect and anti-noise performance.

$$f_{\mathrm{Mal}} = 2595 \lg \left(1 + \frac{f}{700} \right)$$

Principle of DTW algorithm

In time series analysis, dynamic time warping (DTW) is one of the algorithms for measuring similarity between two temporal sequences, which may vary in speed. For instance, similarities in walking could be detected using DTW, even if one person was walking faster than the other, or if there were accelerations and decelerations during the course of an observation. DTW has been applied to temporal sequences of video, audio, and graphics data — indeed, any data that can be turned into a linear sequence can be analyzed with DTW. A well known application has been automatic speech recognition, to cope with different speaking speeds. Other applications include speaker recognition and online signature recognition. Also it is seen that it can be used in partial shape matching application.

In general, DTW is a method that calculates an optimal match between two given sequences (e.g. time series) with certain restriction and rules:

- Every index from the first sequence must be matched with one or more indices from the other sequence, and vice versa
- The first index from the first sequence must be matched with the first index from the other sequence (but it does not have to be its only match)
- The last index from the first sequence must be matched with the last index from the other sequence (but it does not have to be its only match)
- The mapping of the indices from the first sequence to indices from the other sequence must be monotonically increasing, and vice versa, i.e. if $j > i$ are indices from the first sequence, then there must not be two indices $l > k$ in the other sequence, such that index {\displaystyle i}i is matched with index l and index j is matched with index k, and vice versa

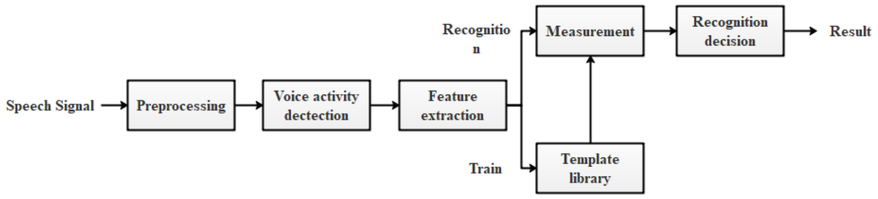
The optimal match is denoted by the match that satisfies all the restrictions and the rules and that has the minimal cost, where the cost is computed as the sum of absolute differences, for each matched pair of indices, between their values.

The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in time series classification. Although DTW measures a distance-like quantity between two given sequences, it doesn't guarantee the triangle inequality to hold.

In addition to a similarity measure between the two sequences, a so called "warping path" is produced, by warping according to this path the two signals may be aligned in time. The signal with an original set of points X(original), Y(original) is transformed to X(warped), Y(warped). This finds applications in genetic sequence and audio synchronisation. In a related technique sequences of varying speed may be averaged using this technique see the average sequence section.

Implementation

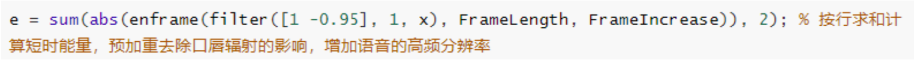
System Flow



Digital speech recognition is to extract the speech features representing the content from a speech, through the analysis and recognition of these features, so as to achieve the purpose of identifying the speech content. Figure shows the flow chart of speech recognition system, which consists of pretreatment, feature extraction, model training, pattern matching and decision strategy. The workflow of digital speech recognition system can be divided into two stages, namely training stage and recognition stage. In the training stage, after preprocessing and feature extraction, the system builds a template or model parameter set for each digital voice through training and learning, and stores these models to form a digital voice model library. In the recognition stage, the test speech is also preprocessed and extracted, compared with the various digital speech templates generated during training, and judged according to certain similarity criteria, so as to determine the corresponding number.

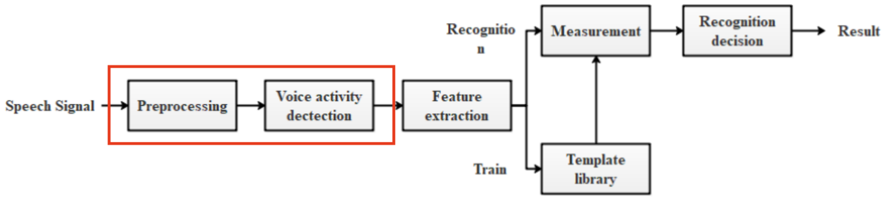
Endpoint detection of speech signal is a basic step in speech recognition. It is the basis of feature training and recognition. The so-called endpoint detection is to exclude the silent segment from the voice signal by locating the beginning and end points of various segments in the voice signal. In the early stage, the main basis of endpoint detection is energy, amplitude and zero-crossing rate. But the effect is often not obvious. In the 1960s, Itakura, a Japanese scholar, proposed a dynamic time warping algorithm. The idea of the algorithm is to extend or shorten the unknowns evenly until they are consistent with the length of the reference mode. In this process, the time axis of unknown words is distorted or bent unevenly, so that their characteristics are aligned with those of the model. Continuous speech recognition is still the mainstream method. At the same time, many improved DTW algorithms have been proposed in the isolated word recognition system.

1) Preprocessing module Before calculating speech parameters, a pre-emphasis filter is usually used, that is:



```
e = sum(abs(enframe(filter([1 -0.95], 1, x), FrameLength, FrameIncrease)), 2); % 按行求和计算短时能量，预加重去除口唇辐射的影响，增加语音的高频分辨率
```

2) Endpoint Detection Module

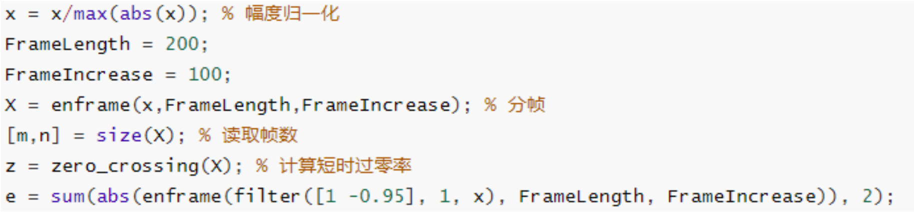


Details and analysis: The endpoint detection of the whole speech signal can be divided into four segments: mute, transition, voice and end. A variable status is used to represent the current state in the program. In the silent section, if the energy or zero-crossing rate exceeds the low threshold, it should start marking the starting point and enter the transition section. In the transition section, because the parameters are relatively small, it is not sure whether they are in the real voice segment, so as long as the values of both parameters fall below the low threshold, the current state will be restored to the silent state. If either of the two parameters in the transition segment exceeds the high threshold, it can be assured that it will enter the speech segment.

Voice activity detection:

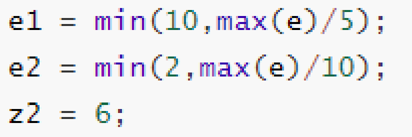
Speech signal is divided into four segments(states): silent segment, transition segment, voice segment and end segment.

- Two important measured value: short-time energy and zero-crossing rate:



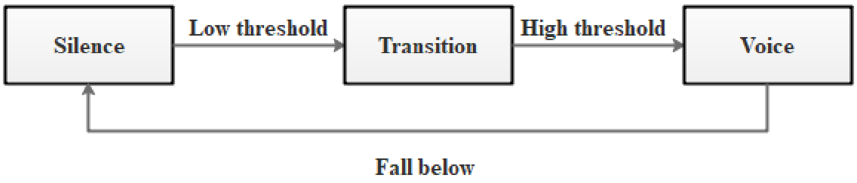
```
x = x/max(abs(x)); % 幅度归一化
FrameLength = 200;
FrameIncrease = 100;
X = enframe(x,FrameLength,FrameIncrease); % 分帧
[m,n] = size(X); % 读取帧数
z = zero_crossing(X); % 计算短时过零率
e = sum(abs(enframe(filter([1 -0.95], 1, x), FrameLength, FrameIncrease)), 2);
```

- Set the low and high threshold:



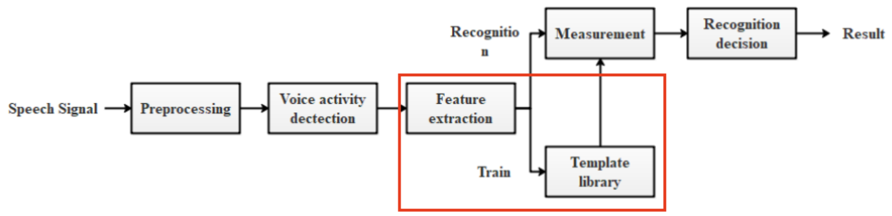
```
e1 = min(10,max(e)/5);
e2 = min(2,max(e)/10);
z2 = 6;
```

- The default state is silent state. In the silent state, if the energy or zero-crossing rate exceeds the low threshold, it should record the starting point and enter the transition section.In the transition state, because the energy and zero-crossing parameters are low, we can't be sure if the current signal is really in the voice segment.If the energy parameters in the transition section exceed the high threshold, the signal is in the voice segment. But as long as the values of both parameters fall below the low threshold, the current state will be restored to the silent state.



```
for i = 1:m
    switch(state)
        case {0,1}
            if e(i)>e1
                start_point = i-count1;
                state = 2;
                count2 = 0;
                count1 = 0;
            elseif e(i)>e2||z(i)>z2
                state = 1;
                count1 = count1 + 1;
            else
                state = 0;
                count1 = 0;
            end
```

3) Feature extraction module

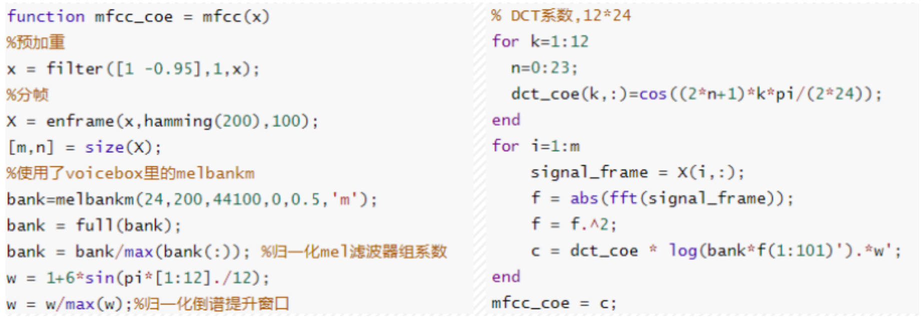


Details and analysis: After endpoint detection, the speech signal needs to extract digital speech features, and the feature parameters need to meet the criteria of minimizing the distance between classes and maximizing the distance between classes. Feature parameter extraction is the basis of digital speech recognition, and its selection directly affects the performance of digital recognition. Referring to the selection of digital speech features in relevant literature, Mel cepstrum coefficients are used as the feature parameters of digital speech recognition in this system. The feature extraction module transforms the pre-processed speech sample database into a feature vector database and saves it for subsequent speech model recognition.

Our extracted parameters is Mel frequency cepstrum coefficient(MFCC). We can get Mel frequency cepstrum coefficient by obtaining Spectrogram, passing through Mel filter , logarithm and IDFT/DCT. And as our sample is small, we choose to add some noise to the sample to increase the difficulty of our training.

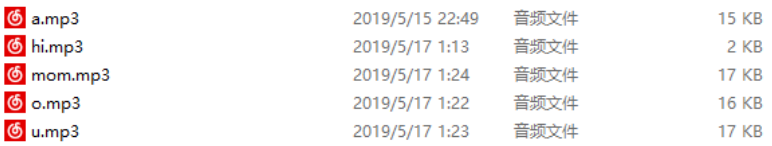
- The way get Mel frequency cepstrum coefficient

1. Obtaining Spectrogram
2. Pass through Mel filter
3. Logarithm
4. IDFT/DCT

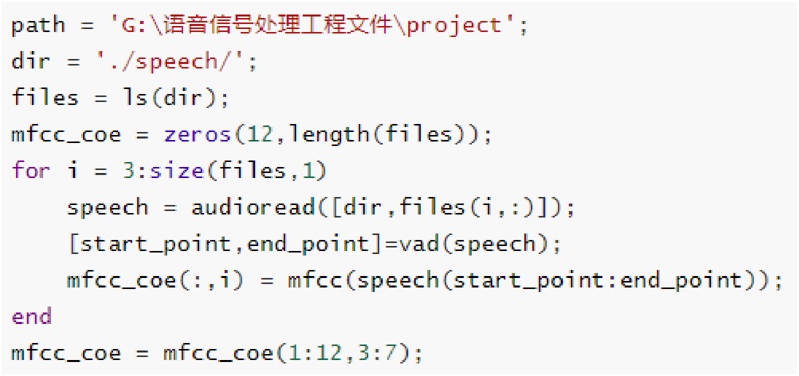


```
function mfcc_coe = mfcc(x)
%预加重
x = filter([1 -0.95],1,x);
%分帧
X = enframe(x,hamming(200),100);
[m,n] = size(X);
%使用了voicebox里的melbankm
bank=melbankm(24,200,44100,0,0.5,'m');
bank = full(bank);
bank = bank/max(bank(:)); %归一化mel滤波器组系数
w = 1+6*sin(pi*[1:12]./12);
w = w/max(w);%归一化倒谱提升窗口
% DCT系数,12*24
for k=1:12
  n=0:23;
  dct_coe(k,:)=cos((2*n+1)*k*pi/(2*24));
end
for i=1:m
    signal_frame = X(i,:);
    f = abs(fft(signal_frame));
    f = f.^2;
    c = dct_coe * log(bank*f(1:101)').*w';
end
mfcc_coe = c;
```

- Extract the feature of library: Read the files in the voice library and extract MFCC.

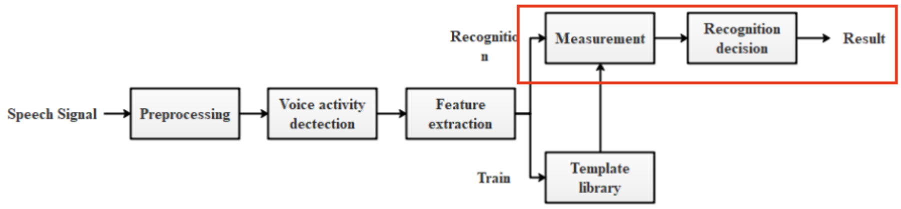


a.mp3	2019/5/15 22:49	音频文件	15 KB
hi.mp3	2019/5/17 1:13	音频文件	2 KB
mom.mp3	2019/5/17 1:24	音频文件	17 KB
o.mp3	2019/5/17 1:22	音频文件	16 KB
u.mp3	2019/5/17 1:23	音频文件	17 KB

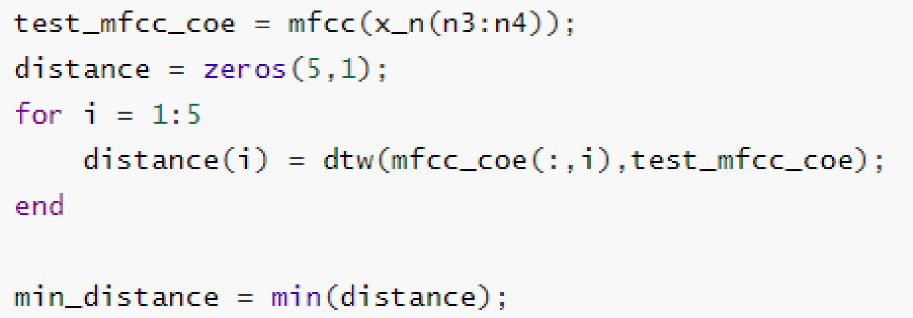


```
path = 'G:\语音信号处理工程文件\project';
dir = './speech/';
files = ls(dir);
mfcc_coe = zeros(12,length(files));
for i = 3:size(files,1)
    speech = audioread([dir,files(i,:)]);
    [start_point,end_point]=vad(speech);
    mfcc_coe(:,i) = mfcc(speech(start_point:end_point));
end
mfcc_coe = mfcc_coe(1:12,3:7);
```

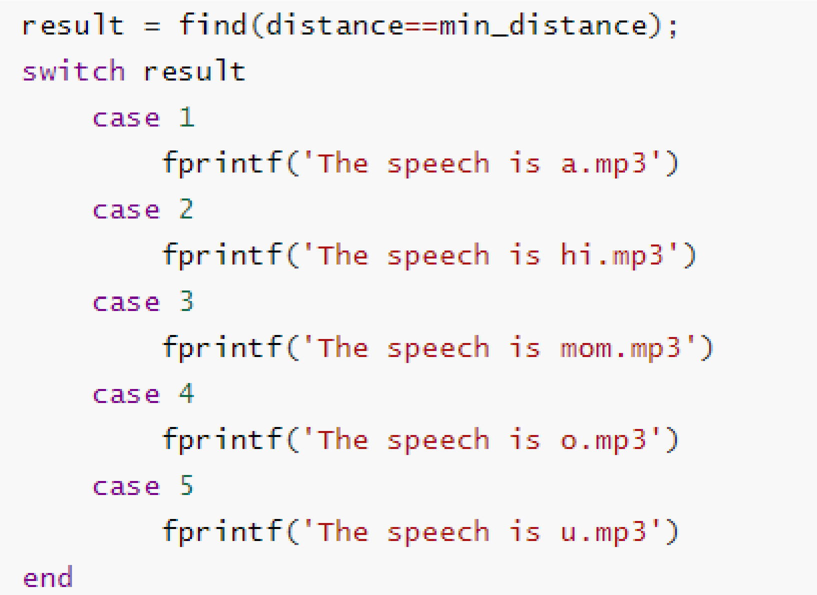
4) Recognition module



Details and analysis: The recognition module mainly completes the judgment of the digital category of speech, and displays the recognition results in the system view area or stores them in the form of text files. The features used in the recognition process are consistent with the types of features in the corresponding training model, and the feature parameters of the speech samples to be recognized are processed the same as those of the training speech samples.



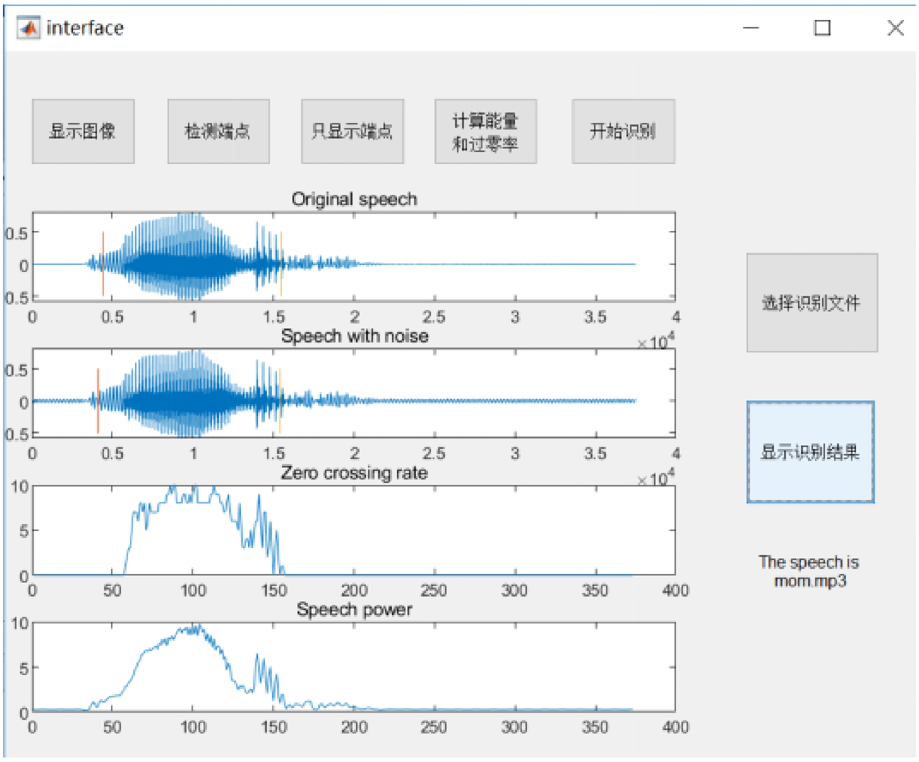
```
test_mfcc_coe = mfcc(x_n(n3:n4));
distance = zeros(5,1);
for i = 1:5
    distance(i) = dtw(mfcc_coe(:,i),test_mfcc_coe);
end

min_distance = min(distance);
```



```
result = find(distance==min_distance);
switch result
    case 1
        fprintf('The speech is a.mp3')
    case 2
        fprintf('The speech is hi.mp3')
    case 3
        fprintf('The speech is mom.mp3')
    case 4
        fprintf('The speech is o.mp3')
    case 5
        fprintf('The speech is u.mp3')
end
```

Achievement and GUI

Detailed process analysis and result analysis methods have been mentioned above. The following figure is the system interaction interface that we ultimately realized.



The speech is mom.mp3

Future Improvement

Since we have fewer samples at the present stage, we can consider adding more samples for training in the future. At the same time, we can also realize real-time speech signal recognition according to the characteristics of different isolated words, which is the process of simultaneous interpretation. In addition, we can also choose other methods, such as DTW, HMM, etc. for speech recognition, compare the recognition effect of different methods, and choose one or more suitable recognition methods.