HW5 report

This assignment is to give two sets of voice comparisons, which are the spoken voices of the same person 0-9 vocal. We need to use mel-spectrogram and plp-rasta to extract voice features. Then, the DTW algorithm is used to train and recognize the voice. After the test group voice is learned, it needs to be able to recognize the comparison group voice, and compare the advantages and disadvantages of the two feature extraction methods.

This time I submitted a work with many files uploaded, of which train.m is used to train the test set. The dtwtest.m file is used to import train data and complete the final alignment. vad.m is a speech preprocessing file used to preprocess the original speech and detect the start and end points of the speech signal. PLP.m and mel.m are the feature extraction of two methods. These two are based on matlab and the encapsulation functions given by the homework and have been modified to some extent. All the remaining files are matlab buildin functions. Due to some problems with my matlab, these files in voicebox and rastaplp cannot be recognized normally after downloading and installing, so I copied them and put them directly in the project.

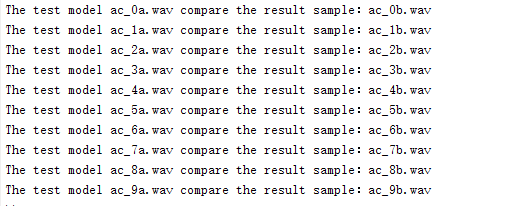
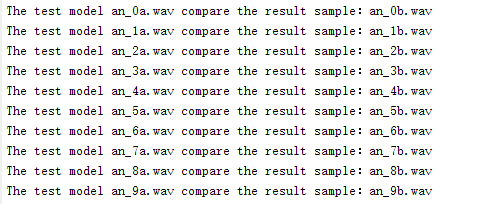
The first is the preprocessing part, what I am doing here is the standard preprocessing of the dtw algorithm. Since the speech signal is a typical non-stationary signal, coupled with respiratory airflow, external noise, current interference, etc., the speech signal cannot be directly used to extract features, but requires preprocessing in the early stage. The preprocessing process includes pre-filtering, sampling and quantization, framing, windowing, pre-emphasis, endpoint detection, etc. The preprocessed speech data can be used for feature parameter extraction.

Then is the feature extraction part. There are two cases here. The first is the feature extraction of mel-spectrogram, and the other is the feature extraction of plp-rasta.

In the feature extraction of mel-spectrogram, the same method as the built-in function of matlab is basically used here, but there is an additional cepstrum processing. The basic idea is to perform pre-emphasis, framing and windowing on the speech, and then obtain the corresponding spectrum through FFT for a short-term analysis window. Then perform mel feature extraction on the obtained spectrum, and obtain the mel spectrum through the filter. Finally, perform inverse transformation with dct discrete cosine to obtain the final spectral coefficient, which is the speech feature.

The other part is the plp-rasta method, the specific principle will not be introduced here. Here I basically directly call the encapsulation function of rastaplp. The only extra part is to calculate the 1st derivative and the 2nd derivative, and then combine them with the features calculated by the function. That is, the same dynamic analysis merge as in the mel feature calculation.

Next comes the training and recognition phase. In the training phase, after the feature parameters are processed to a certain extent, a model is obtained for each entry and saved as a template library. In the recognition stage, the speech signal passes through the same channel to obtain speech parameters, generates a test template, matches with the reference template, and takes the reference template with the highest matching score as the recognition result. Here, according to the requirements of the job, the dtw algorithm is used for processing. dtw mainly uses the time warping function W(n) that satisfies certain conditions to describe the time correspondence between the test template and the reference template, and solves the warping function corresponding to the minimum cumulative distance when the two templates are matched. Here, the function in voicebox is directly called to compare dtw. compute DTW distance between the spectrogram of that test utterance and all training utterances and choose the class that has the smallest distance.

I saved the screenshots of the comparison results obtained by the two methods in the result folder. A few screenshots are shown here. It can be seen that both methods are very good, almost all of them recognize the speech without error. 

Due to the relatively small amount of data, the result I got here is that almost both methods can recognize the word speech very well. If it is a large amount of data, it should have a more intuitive feeling. Here we can only analyze it based on the theory of two feature extraction methods. The first is mel-spectrogram, which is widely used in HMM-based speech recognition systems as acoustic features. Relative spectral features have less correlation, which means that it is easier to build models than spectral features. Its representation is very compact, and the performance of speech recognition is better than that of filter banks or spectral features. The disadvantage is that it has bad robust against noise. The second is plp feature extraction, since it uses equal loudness pre-emphasis and cube root compression (as a result of perception), rather than the logarithmic compression used by Mel, using a linear predictive autoregressive model to obtain cepstral coefficients. So it has better speech recognition accuracy and better noise robustness. But in terms of performance, both are excellent.