Real-time Digit Speech Recognition by HMM

Taoran Lv, Chao Zhang, Dan Zhu

Abstract—The project we will do is real-time, small vocabulary, user-dependent digit recognition. Our goal is to recognize a series of digit which the speaker says through the microphone, and export the digits one by one on the screen immediately. For the speech recognition algorithm, we use the Hidden Markov Model. However, this problem is quite different from the homework 4 we did in the speech class. The "real time recognition" and "a series of digits" have been emphasized in out project. This project could be used in many areas after a little amelioration for the specific function. For instance, handicapped people may use it to record the phone number; it can also be used as digit dictation software. So it makes sense to work on this project.

. System Overview

Many researches have been done on this area, such as different implementation of a hidden Markov model [1], a large vocabulary continuous speech recognition [2], recognition on different languages [3], and also some new approach for recognition using DSP. Based on these previous works and the limited time, we design our system as follow.

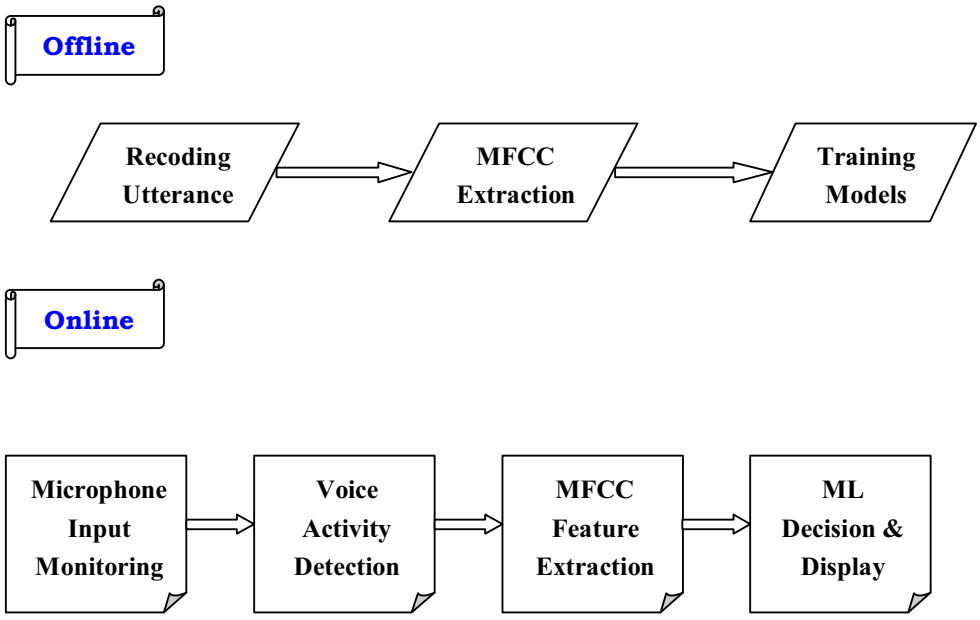

Fig.1 System overview

Fig.1 is the flow chart of our project. Refer to Fig.1, our work can be divided into two parts. One is the offline part. This part is to record utterance and do training to get a good model to be used in the online part. The other is the online part. This part is the main body of our system and it is much more complicated than the offline part. First, let the program monitor the analog input through the microphone. Data Acquisition (Section Ⅱ) doesn't save any data until the analog

input goes up to the threshold. Then, use Voice Active Detection (Section Ⅲ) to get the clean part of voice out. After that, use MFCC to extract the feature, and use Hidden Markov Model (Section Ⅳ) to decide the most likelihood one as the correct result. Finally, show the result on the screen. To emphasize, during the detection, the program is always monitoring the input through microphone in case of loss of data.

Ⅱ. Data Acquisition

Due to the "real time" characteristic of our project, we have to enable Matlab to interface with the sound card of our laptop. We need a gateway to the hardware's functionality and enable us to control the behavior of acquisition. Also, we need to acquire signals from sound card, acquire data to memory, read data into the workspace, and preview the most recently acquired data. All of these can be achieved using data acquisition toolbox of Matlab.

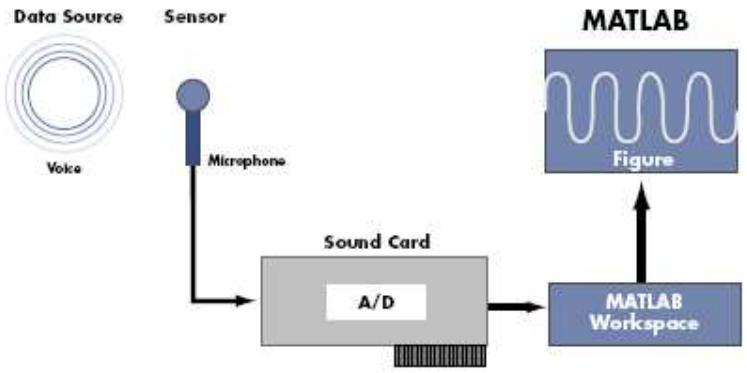

Fig.2 A setup for data acquisition and analysis session.

Also, all the parameters should be set properly to make the data acquisition part work well. For instance, we have to set a threshold to trigger the data acquisition; otherwise, the program would constantly save the analog input in the memory careless about it is our voice or noise, and finally result in a dead loop or memory exceed.

Ⅲ. Voice Activity Detection

Voice activity detection or voice activity detector is an algorithm used in speech processing wherein, the presence or absence of human speech is detected from the audio samples. The main uses of VAD are in speech coding and speech recognition. A VAD may not just indicate the presence or absence of speech, but also whether the speech is voiced or unvoiced, sustained or early, etc.

Voice activity detection is the process of separating conversational speech and silence. The primary function of a voice activity detector is to provide an indication of speech presence in order to facilitate speech processing as well as possibly providing delimiters for the beginning and end of a speech segment. We use VAD to acquire a speech segment as "clean" as possible, which is not only remove some noise but also keep the unvoiced part in the data. We set two constant thresholds in VAD to detect the voiced and unvoiced part, so before the detection, we normalize the data we acquire from data acquisition part. While the normalization makes SNR of speech

segment decreasing to some extent, and affect the final result of the detection.

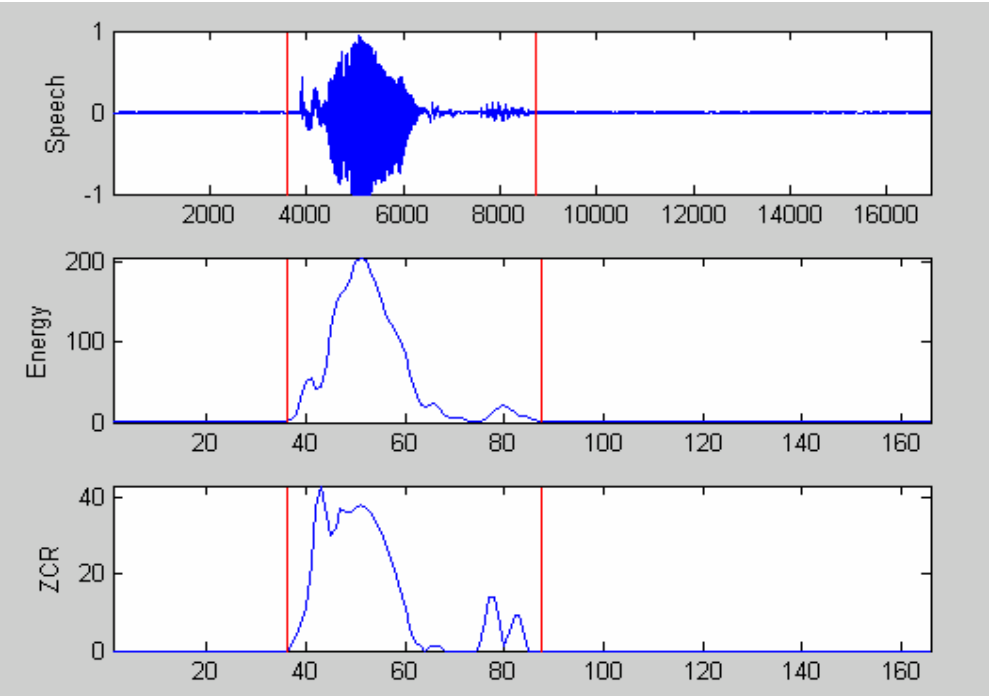

Fig.3 Example of VAD result, test word is ”test”

Fig.3 is an example for our VAD part. The word used in the example is “test”, which contains voiced and unvoiced parts. The segment between the two red color lines is the human voice part. From the figure, we can see that our program works perfectly well in voice activity detection when the threshold set properly.

Ⅳ. Hidden Markov Model

This part is mostly the same as homework 4. We use HMM algorithm to train our utterances. We use five states to model each word, and just use one Gaussian to model for each state. At last, we use maximum likelihood decision to get the result. To get a higher accurate percentage and a better performance, we enlarge the training size. We record 50 utterances for each digit.

Ⅴ. Result and Analysis

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Correct Percent	**>95%**	**>95%**	**>95%**	**>95%**	**>95%**	**>95%**	**>80%**	**>95%**	**>80%**	**>95%**
Most likely mistakes						**9**	8		6	5

From this chart, we can see that we have already got a nearly perfect recognition result. The only possible problem is the utterance 6, which could be recognized as 8 sometimes, though rarely happens. It is a long way to boost the accuracy from 50% to >99%, we made a great effort to try different kinds of ways to do so.

There are two main factors affect the recognition result, which we present as follows respectively:

The first one is the process of training, which can be illustrated from three aspects.

a) The environment during training

The pre-trained model for test might be inaccurate; the best result is got when we do the test in exactly the same room as we record the training data. This aspect even determines the result when we use a small amount of training data. (such as 10 utterances for one digit) The approach to solve this problem is online training, but since it is an incredible time-consuming process, it is not applicable in our real-time recognition system.

b) The size of training database

The increase of the size of training database is a good approach to better our result; we can get a better model in this way obviously. As a matter of fact, the accuracy of recognition rises up by 30 percents when we extending the database by five times, from 10 utterances per digit to 50 utterances. If we want to make the system person-independent, we can just keep extending the database of each person, than combine them.

c） Speaking way during training

This also effect the result a little bit, although not as much as the previous two aspects. since we cannot make sure people say the same word always in the same way, it is necessary to record different kinds of utterances for one digit, to make sure the test utterance can still be recognized even it is spoken in a weird way.

The second main factor is the environment of testing

This is a noise sensitive system, the performance of recognition by HMM dramatically affected by the presence of noise during test. If the noise is too high, we cannot get an accurate speech segment during VAD, let alone the right MFCC and log likelihood. What we can do to improve is filtering or noise estimation, problem will be solved as long as the SNR is high enough to let the VAD process work well.

Reference:

[1] Mitchell, C.D.Helzerman, R.A. Jamieson, L.H. Harper, M.P.Sch. of Electr. Eng., Purdue Univ., W. Lafayette, IN, USA; "A parallel implementation of a hidden Markov model with duration modeling for speech recognition " Parallel and Distributed Processing, 1993. Proceedings of the Fifth IEEE Symposium on pp.298-306, Dec. 1993.

[2] T.Starner and A.Pentland, "Real-Time American Sign Language Recognition From Video Using Hidden Markov Models," Technical Report 375, MIT Media Lab, Perceptual Computing Group, 1995. Earlier version appeared *ISCV'95.*

[3] D'Orta, P.Ferretti, M.Martelli, A.Melecrinis, S.Scarci, S.Volpi, G.IBM Rome Research Center, Roma,Italy, "A speech recognition system for the Italian language", Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87., vol. 12, pp.841-843, Apr. 1987.