Design of Embedded Speech Recognition System

ZHOU Zhi-qiang & YAO Li-ming
Professor
Shanghai University of Engineering Science
College of Mechanical Engineering
China

Abstract

This paper describes the process and the basic principles of speech recognition. The algorithm and signal processing technology of speech recognition are studied in depth. At the same time, the MATLAB software is used to simulate the signal processing method, and verify the feasibility of the method. The system uses STM32F103ZE as the main control chip, and has good expansibility. Speech recognition module uses LD3320 chip. The chip can complete the task of speaker-independent speech recognition.

Keyword: HMM, recognition algorithm, signal processing, STM32, MATLAB, LD3320

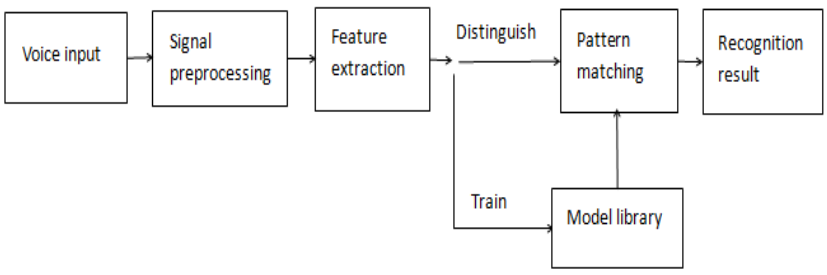
1. Introduction

We use language to realize the interaction between man and computer, mainly including three technologies, namely, speech recognition, natural language understanding and speech synthesis. The primary mission of automatic speech recognition is to complete the transform from the speech to the text. Natural language understanding is to complete the transform from the text to the semantic. Speech synthesis is to output the voice information that users want to output. Today, speech recognition technology has appealed intensively and extensively to increasing number of researchers. Its application will surely bring good social and economic benefits. The embedded speech recognition system has the advantages of small volume, low-power, high reliability and flexible installation. It has quite broad applied foreground in the field of smart appliances and consumer electronics, and become a research focus gradually.

2. The process of speech recognition

After the audio signal is converted into electric signal, the signal needs to be pretreated first. Then the voice model is established based on the characteristics of human language, and the input signal was analyzed. At the same time, we extract the necessary features and set up a template for speech recognition. In the process of recognition, the computer according to the model of speech recognition, compare voice templates and characteristics of the input speech signal. According to a certain search and match mechanism to obtain a series of the best template that matching with the input signal. Then, according to the definition of the template, we can obtain the recognition results by searching (LIU Yao-he, 2008).

Figure1. Speech recognition process



3. Speech recognition algorithm

3.1 HMM profile

At present, the most popular methods of speech recognition have dynamic time warping, hidden Markoff model and artificial neural network method. Among them, the HMM has become the mainstream technology of speech recognition, and widely use in analyzing time series data(Akira Hayashi,2013).Now, most of the speech recognition system of speaker-independent is based on HMM. The HMM describes a probabilistic process that generates an "output" (in our setting a sequence of phonemes). We will construct an HMM for every word, with the goal that the model for w will generate w' as output with higher probability the more perceptually similar w and w' are (Siniscalchi, 2013).HMM is a five tuple (Ωx, Ωo, π, A, B), it usually can be expressed as $\lambda=(\pi,A,B)$

.

$\Omega_x=\{q_1,q_2,\ldots\ldots,q_N\}$, it represents a finite set of hidden states.

$\Omega_o=\{p_1,p_2,\ldots\ldots,p_M\}$, it represents a finite set of observed states.

$\pi=\{\pi_i\}$, it represents the probability of the initial state.

$A=\{a_{ij}\}$, it's a state transfer matrix. $a_{ij}=P(x_{t+1}=q_j|x_t=q_i)$, it represents the probability of a hidden state to another.

$B=\{b_{ik}\}$, it is called the confusion matrix. $b_{ik}=P(O_t=p_k|x_t=q_i)$, it indicates that the probability of anobservation state is obtained when the condition of a hidden state is given.

In the process of speech recognition, the voice of the people is regarded as observable state, and the sound output within the component is considered as a hidden state. The observation state is generated by the speech process, and the observable state is very similar to the hidden state.

3.2 HMM assessment problem: Forward algorithm

In the process of speech recognition, we will use a large number of HMM, each model corresponds to a word. The sound signal is identified by searching for a HMM which is most likely to produce a sequence of observations formed by sound (Marc Lassonde, 2013). We can solve this problem by the Forward algorithm. In the process of calculating, we use recursion to avoid exhaustive calculation of all paths. The computation complexity can be reduced by using the local probability. Under the state j, t time local probability formula is as follows:

$$\alpha_t(i) = P(\text{observations state}|\text{hidden state j}) \times P(\text{T moment all the path to the j state}) \quad (1)$$

When t=1, there is no path to the current state, and the local probability is equal to the product of the initial probability of the current state and the relevant observation probability.

$$\alpha_1(j) = \pi(j) \times b_{jk_1} \quad (2)$$

When t>1, the number of paths required to calculate the local probability rises exponentially as the observed state increases. But the local probability of the T-1 time contains all the probabilities with the former path to this state. So, the local probability of t time is obtained by the local probability of the previous time (Nishanth Ulhas Nair, 2010).The calculation formula is as follows:

$$\alpha_t(j) = [\sum_{i=1}^{N} \alpha_{t-1}(i)a_{ij}]b_{jk_t} \quad (3)$$

From the formula can see that the probability of the observation sequence is calculated recursively after given HMM, and the probability of the sequence will be equal to the sum of all the local probabilities in the T time.

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad (4)$$

3.3 HMM decoding problem: Viterbi algorithm

Decoding is finding an optimal sequence of hidden states after given the observation sequence and the HMM. So that the system can respond to the sound signal, and output the corresponding processing results. The Viterbi algorithm is adopted to decode.

In the Viterbi algorithm, we also define a local probability δ. Different from the local probability of the forward algorithm, it indicates that the most probable path to a state in the t time, rather than the sum of the all paths'

probabilities. When t=1, the most likely path to reach a state does not exist, and the partial probability is equal to the product of initial probability and the corresponding observation probability.

$$\delta_1(i) = \pi(i) \times b_{ik_1} \qquad (5)$$

When t>1, the most probable path to the next state is determined by the following two formulas. At the same time, it records how to get to the next state.

$$\begin{aligned} \delta_t(i) &= \max_j\{\delta_{t-1}(j) \cdot a_{ji} \cdot b_{ik_t}\} \\ \emptyset_t(i) &= \arg\max_j[\delta_{t-1}(j) \cdot a_{ji}] \end{aligned} \qquad (6)$$

Using the formula

$$i_T = \arg\max [\delta_T(i)] \qquad (7)$$

to determine the most likely hidden state when the system is completed. For t<T, according to the formula

$$i_t = \emptyset_{t+1}(i_{t+1}) \qquad (8)$$

We can backtrack path in the entire network according to the most probable state path.

3.4 HMM learning problem: Baum-Welch algorithm

Learning problem is to find an optimal HMM for a given observation sequence and a hidden state sequence, which produces the maximum probability of observation state. Usually, we use Baum-Welch algorithm to solve the learning problem. This algorithm estimates the initial parameters of HMM at first. Then reestimate is based on the observation sequence and the initial model, so we can get a new set of parameters and a new model (Tiberiu Chis, 2015). If probability of the new model's observation sequence is larger than the original model, which indicates that the revaluation model is better. Repeat this process until the probability is no longer increase obviously.

4. The speech signal processing

4.1 Speech signal preprocessing

The speech is a time-varying process; its processing relies on the concept of short-time analysis. Speech signal spectral components are generally concentrated in the 300-3400Hz, so that we should use a band pass filter to intercept the signal from the range. Then the speech signal is sampled, and the analog signal is extracted in the time domain. We should ensure that the sampling frequency must be greater than twice the maximum frequency of the signal. Next the signal is segmented into frames, the speech signal is decomposed into overlapping equal intervals called the frames within which the properties of the signal vary weakly and can be regarded as quasi stationary. Frame is formed by multiplying the signal with window. We often make use of the Hanning windows that ensure low level of the side lobes of the window frequency response (Kolokolov, 2002).Hanning window make the amplitude of the original data changing. In order to maintain the original amplitude, we will need to ensure that have a 1/2 of the overlap between each frame data. We use MATLAB software to simulate the process; the simulation result is as follows.

Figure 2.Time domain and frequency domain of the original signal

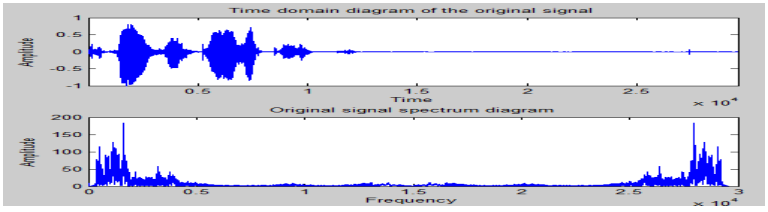


Figure 3.Graphical after pretreatment processing

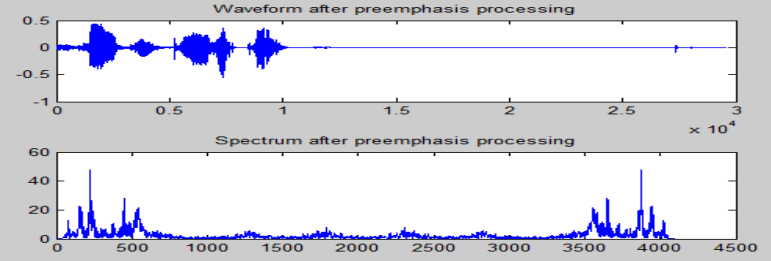


Figure 4.Graphical after frame processing

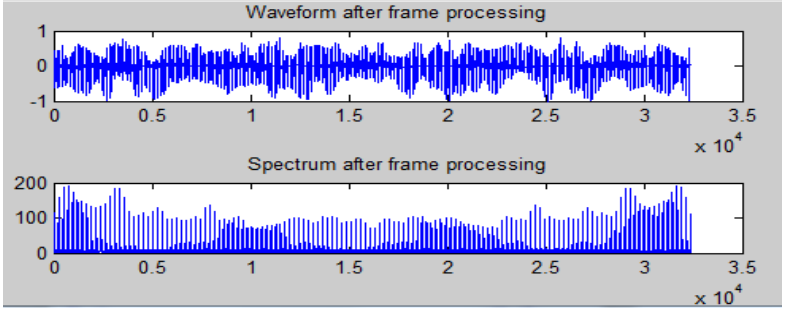
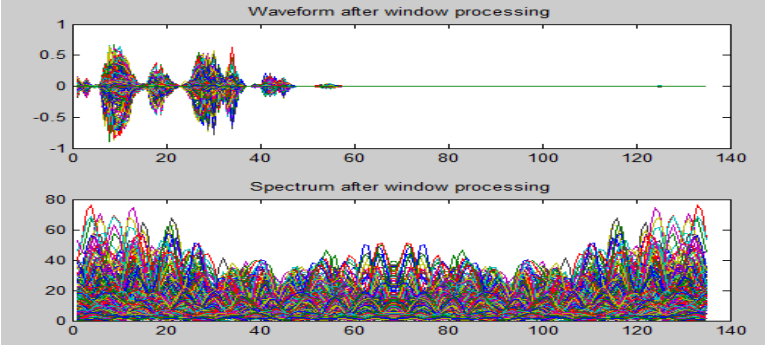


Figure 5.Graphical after windowed processing



4.2 Endpoint detection

Speech endpoint detection is very important in speech recognition; it can distinguish the voice signal and non-voice signals from the noise environment, and then determines the start and endpoints of the operation. The detection is an important preparatory work of speech recognition, the accuracy of the detection direct impact to the feature extraction and the final recognition results (LI Jie, 2012). The experimental statistics show that the deviation of the starting point and the ending point has a significant influence on the final identification of speech recognition. The migration within the 30ms can make the precision drop 2%, when more than 90ms, the impact reached 30% (ZHAO Li, 2005). Under higher signal-to-noise ratio, we can use the energy to distinguish whether there is no speech signal. After the speech signal is segmented into frames and windowed processing, we can get the Nth frame voice signal $x_n(m)$.

$$x_n(m) = w(m)x(n+m) \quad 0 \le m \le N-1 \qquad (9)$$

N is the frame length, so the short-time average energy of $x_n(m)$:

$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \qquad (10)$$

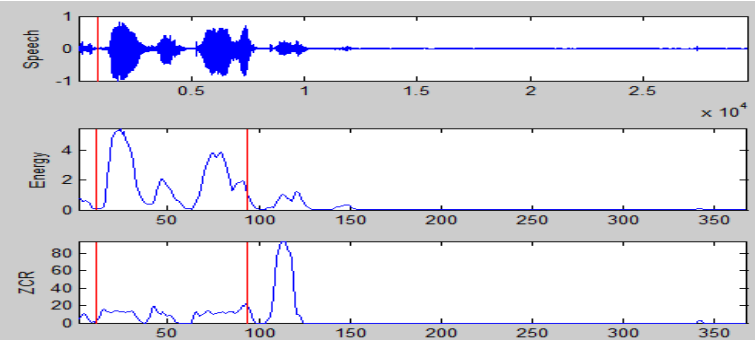
In the initial stage of speech usually has a voiceless of weak energy, it is hard to separate it from the silent zone by energy only. But the voiceless zero-crossing rate (ZCR) is higher than silent zone. So it can be accurately judged by the combination of energy and zero-crossing rate.

The zero-crossing rate is changing the number of speech signal amplitude in a second. After determining the starting point N1 and the end point N2 by energy, the direction of the N1-25 frame and N2+25 frames is searched, and the zero-crossing rate of each frame is compared. If more than three consecutive frames to maintain zero-crossing rate is greater than zero-crossing rate threshold, the starting point is the first frame of meet the conditions, the end point is the last frame. Otherwise, the starting and stopping point is N1 and N2. The calculation formula of the zero crossing rate of the speech signal in the N frame is as follows:

$$Z_n = \frac{1}{2}[\sum_{m=n}^{n+N-1} |sgn(x_n(m)) - sgn(x_n(m-1))|]$$

$$sgn(x_n(m)) = \begin{cases} 1 & x_n(m) \geq 0 \\ -1 & x_n(m) < 0 \end{cases} \qquad (11)$$

Figure6.Endpoint detection simulation result



4.3 Feature parameter extraction

The characteristic parameters of speech recognition mainly include spectral parameters of band pass filter, linear prediction coefficient; linear prediction cepstrum coefficient (LPCC) and Mel-frequency cepstrum coefficient (MFCC).Compared with other parameters, MFCC transforms the linear frequency into Mel frequency. This method outstanding in the information of more efficiency recognition, noise is notably shielded, and MFCC can be used in any case. MFCC is a popular parameterization method that has been widely used in speech technology field especially speech recognition and speaker identification (OoiChia Ai, 2012). The relationship between the linear frequency and the Mel frequency is as follows:

$$f_{Mel} = 2595 \times \log(1 + \frac{f_{Hz}}{700}) \qquad (12)$$

The calculative process of MFCC:

1. After the speech signal is pre-emphasized and windowed, the FFT transform is used to get the spectrum. We are transforming time domain signal into frequency domain signal, and seeking amplitude spectrum.
2. The amplitude spectrum is filtered by the Mel filter banks.
3. We will do logarithmic operation for the output of the filter, and then we can get the MFCC coefficient by the discrete cosine transform.

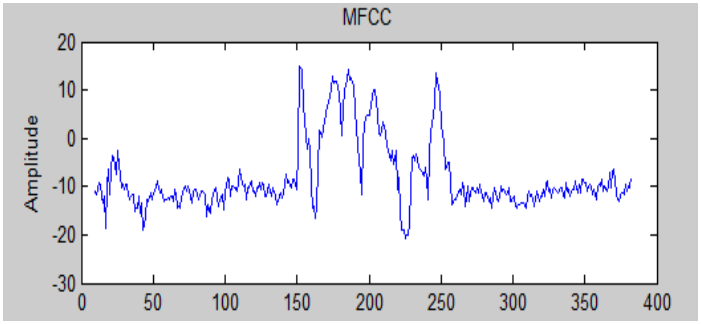
$$C(n) = \sum_{m=0}^{N-1} s(m)\cos[\frac{\pi n(m-0.5)}{M}] \quad (n = 1,2,\ldots\ldots,L) \qquad (13)$$

S (m) is the logarithmic energy of the Mth filter.

4. The standard MFCC only reflects the static characteristics of the speech parameters. The dynamic characteristic of speech can be described by the differential spectral of the static characteristic. Experiments show that the combination of dynamic and static features can effectively improve the recognition property for the system.

When using the MATLAB to simulate the experiment, in order to describe the correlation between each frame, in addition to extracting the MFCC parameters, a first order difference MFCC parameter is introduced in the calculation process (WANG Hua-peng, 2008).MATLAB operating results are as follows:

Figure7.MFCC parameter extraction



5. The hardware design

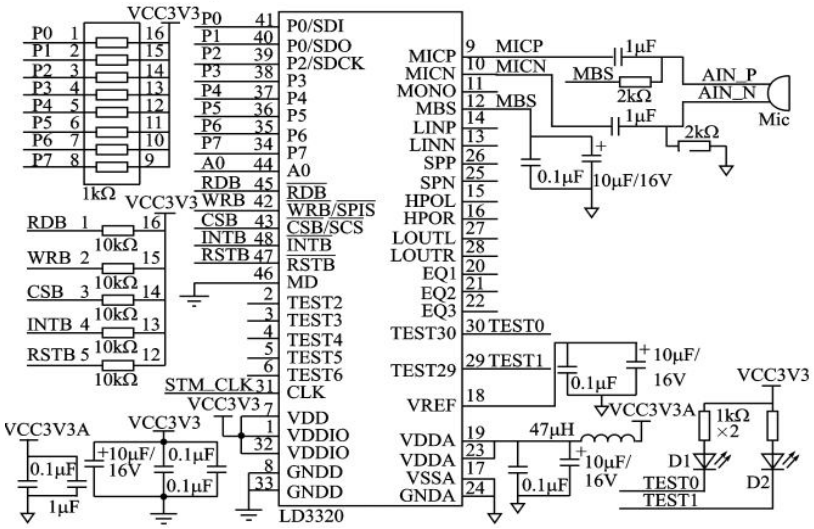
5.1 Master control chip selection

The master controller selects STM32f103ZE control chip of ST Company. The chip is based on the ARM Cortex-M3 32 bit RISC kernel, the operating frequency up to 72 MHz, built-in high speed memory (64KB flash and 20KB SRAM),and has a wealth of enhanced I/O ports and peripherals.

5.2 Speech recognition circuit

Speech recognition module uses LD3320. The chip integrates the speech recognition processor and some external circuits, including AD\ DA converter, microphone interface, audio output interface, etc. It does not need the auxiliary chip, such as flash, ram. It can be directly application in speech recognition and man-machine conversation function. Within the LD3320, there is a highly efficient search engine module of speaker-independent speech recognition and a complete speaker-independent speech recognition feature library, so it is not required to connect the PC machine and download the HMM feature library(ICRoute, 2011).

Figure 8. LD3320 wiring diagram

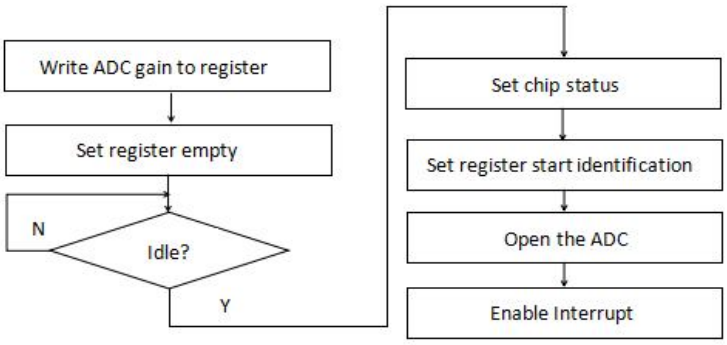


6. The program design

Program design is broadly divided into the following aspects:

1) Initialization, this process mainly completes the reset, the mode setting, the clock frequency setting, the FIFO setting and so on.
2) Write identification list, each identification entry corresponds to a specific number. It can contain the same value, but the value is less than 256.
3) Start speech recognition. Figure 9 shows the process of speech recognition, ADC gain is the microphone volume (SU Peng, 2011).

Figure 9. Identification process



4) Response interrupt, when the microphone collects the sound, it will produce the interrupt signal. Interrupt program will analyze the register value to obtain results.

7. Conclusion

This paper studied in detail the HMM model and the speech recognition algorithm. The processing method of speech signal is simulated by using Matlab software and the feasibility of the solution is proved.LD3320 speech recognition module is controlled by STM32. It has a good stability, anti-interference ability and high recognition rate. It has the good use prospect in the embedded speech recognition system.

References

Akira Hayashi, Kazunori Iwata. (2013). Nobuo Suematsu. Marginalized Viterbi algorithm for hierarchical hidden Markov models. Pattern Recognition, 46, 3452–3459.

A.S. Kolokolov. (2002). Signal Preprocessing for Speech Recognition. Automation and Remote Control, 63, 494–501.

ICRoute. (2011). LD3320 Development Manual.[Online] Available: http://www.icroute.com/web cn/ DownLoad.html#M-LD3320

LIU Yao-he, SONG Ting-xin. (2008). Speech recognition and control technology, Beijing: Science Press, (Chapter 1).(in Chinese)

LI Jie, ZHOU Ping, JING Xinxing, DU Zhiran. (2012). Speech Endpoint Detection Method Based on TEO in Noisy Environment. Procedia Engineering, 29, 2655 – 2660.

Marc Lassonde, Ludovic Nagesseur. (2013). Extended forward–backward algorithm. Journal of Mathematical Analysis and Applications, 403, 167–172.

Nishanth Ulhas Nair, T.V.Sreenivas. (2010) Multi-Pattern Viterbi Algorithm for joint decoding of multiple speech patterns. Signal Processing, 90, 3278–3283.

Ooi Chia Ai, M. Hariharan, Sazali Yaacob, Lim Sin Chee. (2012).Classification of speech dysfluencies with MFCC and LPCC features. Expert Systems with Applications, 39, 2157–2165.

Siniscalchi, S. M. ,Yu, D. ,Deng, L. ,Lee, C.-H. (2013). Speech Recognition Using Long-Span Temporal Patterns in a Deep Network Model. Journal of Memory and Language, 20, 88-102.

SU Peng, ZHOU Feng-yu, CHEN lei. (2011). Design of embedded speech recognition module based on STM32. Microcontrollers & Embedded Systems, 2, 42-45. (in Chinese)

Tiberiu Chis, Peter G. Harrison. (2015). Adapting Hidden Markov Models for Online Learning. Electronic Notes in Theoretical Computer Science, 318, 109–127.

WANG Hua-peng, YANG Hong-chen. (2008). Study on extraction methods of voiceprint recognition feature of MFCC. Journal of Chinese People's Public Security University, 01, 28-30.(in Chinese)

ZHAO Li. (2005).Speech Signal Processing, Beijing: Machinery Industry Press, (Chapter 3).(in Chinese)