Embedded Speech Recognition System for Intelligent Robot

Qingyang Hong, Caihong Zhang, Xiaoyang Chen, Yan Chen, *Member*, IEEE

Abstract — Automatic speech recognition (ASR) is a task that requires high computation capability and enough memory, which is difficult to work in the embedded devices. We have successfully developed an embedded speaker-independent speech recognition system. Specifically, we have also successfully designed a hardware module that can be embedded into the toy robot. We evaluated the speech-controlled robot and the recognition performance was quite good.

Index Terms —Automatic speech recognition, speech recognition

I. INTRODUCTION

With the advances of information technology, the next generation of user interface is desired to be more friendly and powerful. As the most natural and expressive means of communication, speech is a suitable choice for the human-computer interaction. Furthermore, as we move from desktop PC to mobile phone, toy and other embedded devices, the user interface becomes smaller in size and the operation may thus be limited by facilities of hardware. Therefore, speech has the potential to provide a direct and flexible interaction for the embedded systems.

Speech control is dependent on the technology of automatic speech recognition (ASR) [1,2,3], the task of which can be speaker dependent or speaker independent. Generally, speaker-independent system is more widely used, since the user is not required to conduct the training. Speech recognition can be also divided into isolated word recognition, connected word recognition or large vocabulary continuous recognition. For the embedded device, it is generally enough to deploy isolated recognition system. And currently one typical application is for the toy which only requires simple and small-vocabulary command controls.

In this paper, we present our work of deploying a speaker-independent speech recognition system for a popular intelligent robot. The organization of this paper is as follows. Section 2 has a brief introduction of automatic speech recognition. In Section 3, we address the designing issues of embedded speech recognition system. Section 4 describes the control mechanism of intelligent robot. Section 5 gives the evaluation results of the designed system. Finally, we give the summarization of our work.

II. OVERVIEW OF SPEECH RECOGNITION

The research of speech recognition originated as early as 1950's [1]. This technology encompasses a wide range of disciplines, which include acoustics, phonetics, pattern recognition, statistics, probability theory, linguistics, and so on.

Speech recognition can be viewed as a pattern recognition task, which includes training and recognition. Generally, speech signal can be viewed as a time sequence and characterized by the powerful hidden Markov model (HMM) [1,2]. Through the feature extraction, the speech signal is transferred into feature vectors and act as observations. In the training procedure, these observations will feed to estimate the model parameters of HMM. These parameters include probability density function for the observations and their corresponding states, transition probability between the states, etc. After the parameter estimation, the trained models can be used for recognition task. The input observations will be recognized as the resulted words and the accuracy can be evaluated. The whole process is illustrated in Fig. 1.

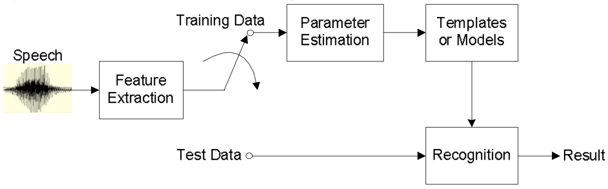

Fig. 1 Block diagram of automatic speech recognition (ASR) system

A. *Hidden Markov Model*

HMM is a probabilistic pattern-matching approach which models a time sequence of speech pattern as the output of a stochastic or random process. HMM has been proven to be one of the most successful statistical modeling methods in the area of speech recognition [2,3].

The Markov generation model is a collection of states connected by transitions. $a_{ij} = P\{q_{t+1} = j \mid q_t = i\}$, $1 \le i, j \le N$ defines the probability of the transition probability from state q_t to q_{t+1} , in which N is the number of states. The output probability distribution can be represented as $\mathbf{B} = \{b_j(\mathbf{o}_t)\}$, in which $b_j(\mathbf{o}_t) = P\{\mathbf{o}_t \mid q_t = j\}$ $1 \le t \le T$, $j = 1,2,\cdots,N$ defines the probability of the observation which is equal to the symbol $\mathbf{o}_t$ in state *j*. The output probability distribution can be either discrete or continuous density. Currently it is

1-4244-1358-3/07/$25.00 ©2007 IEEE

very common to use the continuous density HMM, which has good ability to directly characterize continuous speech in the form of multi-dimensional real-valued feature vectors.

The probability density function (pdf) of continuous HMM is usually in terms of Gaussian functions and has the following form:

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{M} c_{jk} N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})\ 1 \le j \le N, \quad 1 \le t \le T \tag{1}$$

where M is the number of mixture per state, c_{jk} is the mixture coefficient for the k^{th} mixture in state j, and N is the mixture Gaussian density function with the mean vector $\boldsymbol{\mu}_{jk}$ and covariance matrix $\boldsymbol{\Sigma}_{jk}$, which can be written as

$$N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) = \frac{1}{(2\pi)^{D/2}\left|\boldsymbol{\Sigma}_{jk}\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{jk})' \boldsymbol{\Sigma}_{jk}^{-1}(\mathbf{o}_t - \boldsymbol{\mu}_{jk})\right] \tag{2}$$

where prime denotes vector transpose and D is the dimension of the vector $\mathbf{o}_t$. The mixture gains c_{jk} need to satisfy the following constraints,

$$\sum_{k=1}^{M} c_{jk} = 1,\ 1 \le j \le N,\ c_{jk} \ge 0 \tag{3}$$

The covariance matrix can be full or diagonal. It is common to use diagonal matrix for the mixture function, since linear combination of diagonal covariance Gaussians has the same model capability with full matrix and the number of parameters of Gaussian mixtures densities can be reduced. With the diagonal covariance matrix, the Gaussian pdf can be further expressed as:

$$N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) = \frac{1}{(2\pi)^{D/2}\prod_{l=1}^{D}\sigma_{jkl}} \exp\left[-\frac{1}{2}\sum_{l=1}^{D}\frac{(o_{tl} - \mu_{jkl})^2}{\sigma_{jkl}^2}\right] \tag{4}$$

where l is the element index of the vector, o_{tl} is the l^{th} component of the feature vector $\mathbf{o}_t$, μ_{jkl} and σ_{jkl} is the l^{th} mean and covariance in the k^{th} mixture of the j^{th} state, respectively. The denominator is dependent only on the mixture of state. Therefore, we could precompute it and reserve the result for subsequent calculations during the training process.

B. Model Train

For continuous HMM, the re-estimation formulae for the observation density with multiple observation sequence are given as follows

$$c_{jk} = \frac{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)}{\sum_{k=1}^{K}\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)} \tag{5a}$$

$$\boldsymbol{\mu}_{jk} = \frac{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)\mathbf{o}_t^c}{\sum_{c=1}^{C}\sum_{\text{t}=1}^{T_c}\gamma_t^c(j,k)} \tag{5b}$$

$$\boldsymbol{\Sigma}_{jk} = \frac{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)(\mathbf{o}_t^c - \boldsymbol{\mu}_{jk})(\mathbf{o}_t^c - \boldsymbol{\mu}_{jk})'}{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)} \tag{5c}$$

where $\gamma_t^c(j,k)$ is the probability of being in state j at time t with the k^{th} mixture component accounting for $\mathbf{o}_t^c$

C. Probability Calculation

The task of isolated word recognition is to select the word in the vocabulary with the maximum probability given the observation sequence $\mathbf{O}$. Assuming that there are V words in the vocabulary, the maximum selection is defined as

$$\arg\max_v\{P(\lambda_v \mid \mathbf{O})\}, \quad 1 \le v \le V \tag{7}$$

where λ_v represents the v^{th} word model. According to the Bayes' rule, we have

$$P(\lambda_v \mid \mathbf{O}) = \frac{P(\mathbf{O} \mid \lambda_v)P(\lambda_v)}{P(\mathbf{O})} \tag{8}$$

Therefore, assuming that each word in the vocabulary has the same prior probability, the most possible word utterance only depends on the likelihood $P(\mathbf{O} \mid \lambda_v)$.

III. EMBEDDED SPEECH RECOGNITION SYSTEM

To deploy the embedded speech recognition system for small-size toy robot, two issues must be considered. First, a size-suitable hardware module should be designed that can be embedded into the body of robot. Second, the speech model and recognition procedure should be carefully selected or revised to work in the chip.

A. Hardware Module

Our embedded speech recognition system is based on a 16-bit chip, which is widely used in the areas of digital sound process and voice recognition. The memory capacity includes 4K-byte working SRAM plus a 64K-byte flash memory. And there are built-in microphone amplifier and AGC function in this chip.

To let the chip work, the module should be designed in which necessary peripheral circuits are added. Usually, this module is based on encapsulated chip and has a length of 105mm and width of 70mm. This size makes it difficult to be embedded into the body of toy robot.

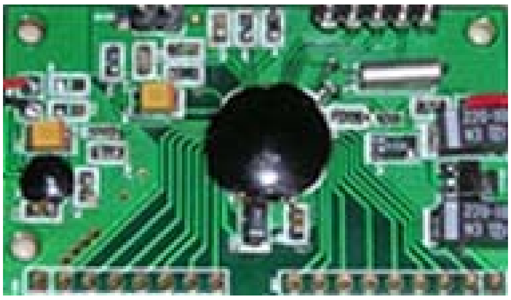

Fig. 2 The hardware module (TS_ASR_MODULE)

Through special hardware technique based on naked chip, the size of our module (Fig. 2) was reduced to the length of 51mm and width of 32mm. This assured that it can be embedded into the body of toy robot. To communicate with the outside objects (e.g. motors in the robot), this module (TS_ASR_MODULE) provides 16 input/output (IO) ports (IOA0~7, IOB0~7).

B. Speech Model

The model unit of speech can be word, phoneme and triphone. While the sub-word unit such as phoneme or triphone has more flexibility for the extension of vocabulary [3], whole word was selected as the model unit for the embedded speech recognition system due to the chip's limited memory capability. Since the recognition system is speaker-independent, the word model units should be trained based on the speech data collected from enough speakers. The range of speakers should cover different sex, age, region, and date.

The memory and processing capability of the chip was limited, and it doesn't support the floating-point and other complex mathematical operations. Therefore, it is difficult to run the training procedure, which has high computation consumption and requires large memory for the training data to build speaker-independent system. To solve this problem, we detach the training procedure and recognition procedure. That is, the speech models are estimated off-line in the PC, and we use them to build the embedded recognition procedure in the chip.

C. Recognition Procedure

To build the chip-based speech recognition procedure, the original PC-based feature extraction and recognition procedure should be transformed line by line from float-point operation to fix-point operation. And the division and exponential operation also need to be replaced with digit-moving and log operation. Specifically, the training and recognition procedure must have the same feature extraction step. This is to assure the transformation from speech data to feature vector is uniform in these two procedures. Otherwise, the estimated models based on trained data can't be used to make correct recognition of test data.

IV. Speech-Controlled Robot

The robot selected for our work is a popular toy, which can do six actions that are controlled by the electrical motors. Based on these six actions, we define the corresponding voice commands: Forward, Backward, Turn Left, Turn Right, Fire and Dance. And they are all uttered in Chinese.

During the recognition process, the response of speech-controlled robot is in real time. Once detecting the voice command, it would make the speech recognition and give a control signal to the corresponding IO port. As shown in Fig. 3, the motor connected with the port will then drive the robot to conduct the required action. For example, if we say the voice command of "Turn Left", the ports of IOB6 and IOB7 will get a signal to drive the motor (Motor_LeftLeg).

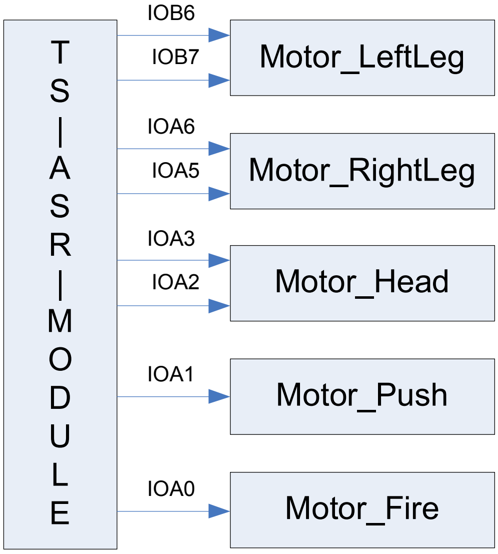

Fig. 3 The IO ports of TS_ASR_MODULE to control the motors

V. Evaluation Results

A. Training Database

To build speaker-independent recognition system, the speech models of voice commands should be optimized based on the enough training database. In our experiment, the training database contains recordings of 40 male speakers and 40 female speakers. For each speaker, there are 6 commands recorded three times under the silent environment. Each recorded data consisted of up to 3s of utterance.

Before using for the training, the exact beginning and ending point of each utterance was adjusted carefully. This step was to discard the silence, noise and other part of the utterance. They were not true data from the recording speaker and might destroy the training performance. After the training, the optimized models were downloaded into the embedded recognition system.

B. *Evaluation Results*

During the evaluation process, the speech-controlled robot would make real-time recognition and carry out the corresponding action. If the corresponding action matched the voice command, it meant that the recognition result was correct.

There were 80 testing speakers (not from those training speakers) selected in total and two evaluations were conducted: 68 speakers in normal environment and 12 speakers in noisy environment. The evaluation results were as follows.

Table 1

Sex	Num	Age	Accuracy
Male	49	20~29	100%
Female	19	20~29	100%
Average			100%

Evaluation in normal environment

As shown in Table 1, our embedded speech recognition system has perfect performance in normal environment. The recognition accuracy even reached 100%, which showed that our speaker-independent system was very robust for those testing speakers.

If the evaluated environment was very noisy, the speech recognition might be destroyed and the accuracy would be reduced distinctly. As shown in Table 2, the average accuracy of our evaluation was 91.7%.

Table 2

Sex	Num	Age	Accuracy
Male	3	10~49	88.9%
Female	9	20~39	94.4%
Average			91.7%

Evaluation in noisy environment

It can be demonstrated that noise was key factor that affect the recognition performance. So far, it is still one of the most difficult tasks in the area of automatic speech recognition. In the future, we'll find more effective means to solve this problem.

VI. SUMMARY

In this work, we have successfully developed an embedded speech recognition system that was speaker independent. The intelligent robot controlled by this system could well recognize the voice commands from the testing speakers and conduct the corresponding actions. 80 speakers were selected to make the evaluation and the results were quite good. In normal environment, the recognition accuracy could reach 100%.

REFERENCES

[1] L.R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, 1993.

[2] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book (for htk version 3.0)*. (htk.eng.cam.ac.uk/prot-docs/HTKBook/ htkbook.html), July 2000.

[3] Xuedong Huang, Alex Acero and Hsiao-Wuen Hon, *Spoken Language Processing*, Prentice-Hall, Inc., 2001.

[4] Yuet-Ming Lam, Man-Wai Mak, and Philip Heng-Wai Leong, "Fixed-Point Implementations of Speech Recognition Systems," *GSPx Conference*, Apr.3, 2003.

[5] Sujay Phadke Rhishikesh Limaye Siddharth Verma and Kavitha Subramanian, "On Design and Implementation of an Embedded Automatic Speech Recognition System," *Proceedings of the 17th International Conference on VLSI Design*, 2004.

[6] D.Wang, J.Liu, Rensheng Liu, Liang Zhang, "Embedded speech recognition system on 8-bit MCU core," *Proceedings of the IEEE International Conference on Acoustics Speech, and Signal Processing (ICASSP)*, 2004.

Qingyang Hong is a tutor of master, born in 1979 in quanzhou, He received his PHD in City University of Hong Kong (2005),majored in speech recognition especially in embedded speaker recognition.
E-mail: qyhong@xmu.edu.cn

Caihong Zhang, female, was born in weihai, in1981. She is a master in Xiamen University and interested in the research of speaker verification.

Chen Yan, Female, was born in April, 1985, in Putian, Fujian, studying as a postgraduate in Xiamen University. She focused on Speaker Recognition.

Xiao-Yang Chen was born in Fujian, in 1981. He is now pursuing his M.E. degree in computer science at Xiamen University, China. His research interests include speech recognition, computer telephony integration etc.