SINGLE-CHIP SPEECH RECOGNITION SYSTEM BASED ON 8051 MICROCONTROLLER CORE

Shi Yuanyuan, Liu Jia and Liu Runsheng
Department of Electronic Engineering, Tsinghua University
100084, Beijing, P.R. China
E-mail: shiyy@hannah.ee.tsinghua.edu.cn

ABSTRACT

This paper describes a single-chip speech recognition system. It contains the speech functions of prompt, playback, speaker-dependent speech recognition, suitable for the voice activated systems in toys, games, consumer electronics, office devices, etc. The chip is designed based on the SOC (System on Chip) philosophy and an 8-bit MCU, RAM, ROM, ADC/DAC, PWM, I/O ports and other peripheral circuits are all embedded in it. Software modules including control/communication, speech coding and speech recognition algorithms are implemented in an 8051 compatible microcontroller core, resulting in the extremely low cost of the chip. The speech recognition adopts the template matching technique. It recognizes up to 20 phrases with an average length of 1 second and the recognition accuracy reaches more than 95% with the background SNR above 10dB. Speech coding uses Continuous Variable Slope Deltamodulation (CVSD) algorithm. The bit rate is 16kbits/s.

1. INTRODUCTION

With the development and maturity of speech compression/decompression and speech recognition, speech is becoming an important form of man-machine interface. This speech interface is increasingly used in office automation, factory automation and home automation devices. By far, most of these systems are large-scale ones. They are proposed to employ the perceptual speech coding method for speech data storage and transfer, or to be capable of giving the synthesized, high quality machine answer. And the speech recognition involved is also some large vocabulary continuous speech recognition system, which is required to obtain a high level of recognition accuracy when responding to the natural, fluent human speech. Therefore these systems have excellent performance. At the same time, they are complicated and used only on the computer platform. But evidently in areas where only simple machine prompt and recognition are required, such as the applications in toys, games, consumer electronics, etc., the low cost systems with a small capacity of prompt, recognition and playback are preferable. Furthermore if the advantages of power consumption, size, cost, integration and reliability are considered, a single chip speech recognition system implementing above speech functions is the best choice and makes wide applications.

As the man-machine interface, the speech recognition system should provide the mutual communication functions. To facilitate its usage for people, it is expected to be able to "speak" some prompts to guide the recognition process and give corresponding information when the recognition is over; to communicate with other devices, it should have several standard I/O ports. Therefore the speech compression/decompression, recognition and I/O functions should be included in the chip. Moreover it is better to design the chip based on the SOC philosophy, that is, to realize the entire system's functionality in a single chip integrating digital and analog units. Then the chip only requires power supply, microphone and speaker for operation, and such a system is of small size, low cost and high reliability.

The speech recognition system described in this paper is a single-chip containing the speech functions of prompt, playback, speaker-dependent phrase training and recognition. The chip integrates an 8-bit microcontroller (MCU) core, on-chip RAM, on-chip ROM, ADC/DAC, bandpass filters, PWM, I/O ports and other peripheral circuits. The MCU core functions as the CPU of the chip, which implements the digital signal processing tasks and the system control. It is object code compatible with the industry standard 8051 microcontroller. Selecting the 8051 core rather than a DSP core is a two-edged knife. The principal benefit is that this speech recognition SOC with the 8051 core costs at least ten times lower than that with a DSP core. So it is preferable for the low price consumer products. But the digital signal processing power of 8051 is so limited that realizing the high performance speech recognition, even small vocabulary, isolated word recognition, on it is a challenge. Not only the speech recognition algorithms must be simplified to fit the source limit of 8051, which likely results in performance degradation, but also some effective methods are necessary to overcome the channel distortion and low signal-to-noise ratio conditions with which the applied systems confront frequently. These considerations are described in the following paragraphs.

The rest of this paper is organized as follows. In Section 2, we describe the heterogeneous architecture of the speech recognition SOC. In Section 3, the algorithmic details of the chip are described. Evaluation results on the test board are given in Section 4. In the last section, we summarize the performance and application fields of our speech recognition SOC.

2. HARDWARE ARCHITECTURE

The proposed speech recognition SOC has a heterogeneous architecture that is composed of an 8-bit MCU core, 512 bytes on-chip RAM, 8k bytes on-chip ROM, 72192×6 bit voice ROM, 10bit ADC/DAC, a PWM (Pulse Wide Modulation), general purpose I/O ports, and other peripheral circuits. The block diagram is shown in Figure 1. The details of each block are described as below.

Contributed Paper

Original manuscript received November 27, 2000

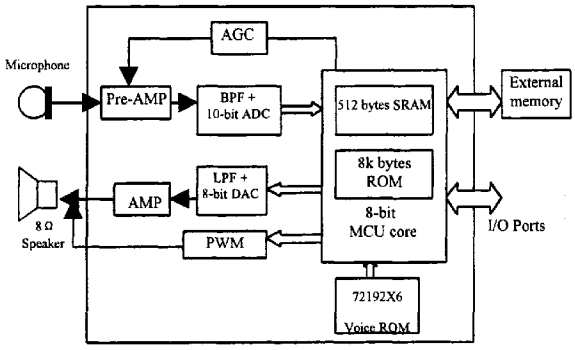

Figure 1. Block diagram of the speech recognition SOC.

- 8-bit MCU core

The 8-bit MCU core is DW8051 MacroCell, the microcontroller core that is object code compatible with the industry standard 8051 microcontroller. It provides increased performance by improving the instruction execution rate by an average factor of 2.5 over the standard 8051 architecture and permitting extended internal RAM space access by multi-using the MOVX instruction. The MCU core performs the speech recognition tasks, including speech feature extraction, speaker-dependent templates training and recognition. It also executes the speech coding/decoding algorithm of Continuous Variable Slope Deltamodulation (CVSD)[1] to restore the prompts, compress and playback the user speech. Furthermore, the MCU core also completes all control and communication tasks.

- Memory spaces

The MCU core can manipulate operands in three address spaces: 64k bytes data memory, 64k bytes program memory and 72192X6 voice data memory. 512 bytes data RAM and 8k bytes program ROM are embedded in the chip. The data and program memory up to 64k bytes can be expanded externally. The on-chip voice ROM is designed specially to store the fixed set of system prompts, which is application driven, obtained with CVSD coding algorithm (16kbits/s). Using a 72192X6 voice ROM, about 27 seconds of speech can be coded and stored in it.

- ADC/DAC

The ADC is a 10-bit channel with the SNR of about 48dB. Before it is the pre-amplifier and the 300~3400Hz bandpass filter. The DAC and PWM are both 8-bit channels. Either of them can be selected as the voice output channel.

- I/O ports

The I/O ports include a serial port, an 8-bit parallel port and a 5-pin serial port.

3. SOFTWARE ALGORITHMS

3.1 Basic Software Modules

The software of speech recognition SOC is composed of three basic modules: control and communication module, CVSD module and speech recognition module. The software algorithm block diagram is showed in Figure 2. The special aspects of the software are described in the following paragraphs.

3.2 Front-end process

The conventional front-end process methods of speech recognition are employed. After the input signal is bandpass pass filtered and sampled at 8k Hz, it is segmented into 24ms frames with 12ms overlap. At the same time, the active speech signal is segmented at the frame level by the energy based endpoint detection. Then 7-order LPCC (Linear Prediction Cepstrum Coefficient)[4] of each pre-emphasized frame is extracted by Schur[2] algorithm. The delta LPCC is also calculated.

During the above process, the most crucial step is endpoint detection. It is well known that the performance of a template based word recognition system is very sensitive to the variations of endpoints. But the simple energy based endpoint detection, which is indispensable in this speech signal driven system, fails frequently to estimate the precise boundaries in the low SNR condition. In order to improve the endpoint detection precision at a little computation cost, the two-stage endpoint estimation is introduced. The time synchronous endpoint detection based on energy and zero-crossing rate is the first stage to estimate the rough active voice boundaries. When the features of the whole voice signal are extracted real-timely, more exact endpoints are estimated at the second stage. First, two energy peaks above an upper threshold are gotten at the front and back ends of the speech signal. Then more exact endpoints are searched from the forward and backward energy peaks respectively, and are determined by the lower threshold. These thresholds are adaptively adjusted according to the volume of the active voice and the background noise. It is found empirically that the upper threshold being 60% of the maximum of the voice's log-energy and the lower one being 7 times the average log-energy of background, which is determined by the experiments. The good detecting results can be obtained using these thresholds. The two-stage endpoint estimation is showed in Figure 3. Then 7-order LPCC feature is extracted. Because 8051 computing ability is very limited, we can only extract the 7-order feature real-timely within the speech limit of the 40MHz MCU core.

In the following paragraphs, several problems existing in the application-oriented products are discussed at the front-end level. It is noted that there exist mainly three factors to degrade the speech recognition performance. One is too small or too large speech volume. The volume being too small results in low SNR and few effective number of quantization bits. If the volume is too loud, the ADC is overloaded. Therefore besides the automatic gain control in the ADC channel, the prompts ask the user to speak softly or to speak loudly if the input speech volume is tested to be too large or too small.

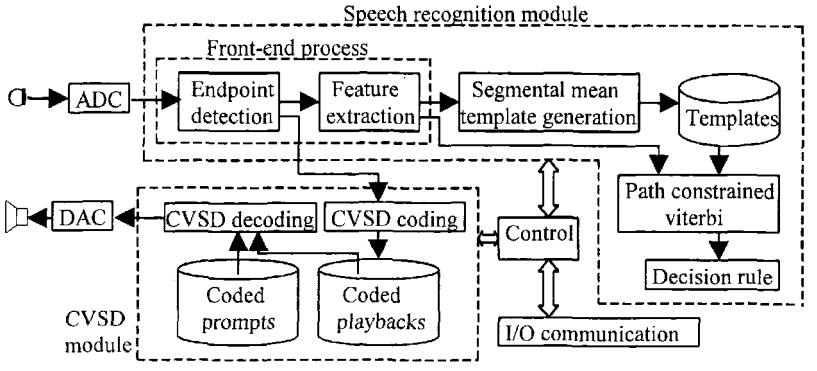

Figure 2. Algorithm block diagram of speech recognition SOC.

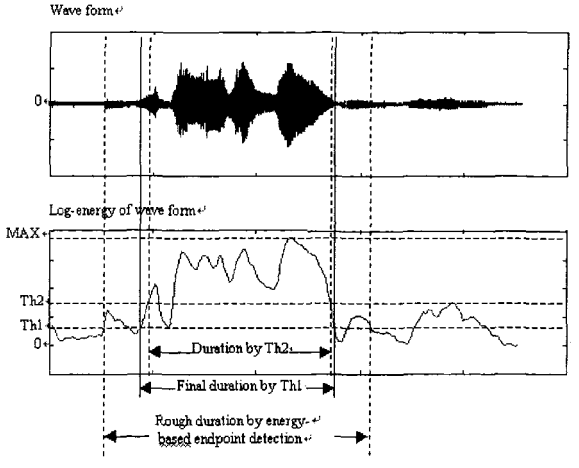

Figure 3. Two stage endpoint detection
MAX: maximum of log-energy of voice
Th2: upper threshold in stage 2, Th2=MAX*ratio
Th1: lower threshold in stage 2.

The second factor is the different distance and direction between the user mouth and the microphone electret. The different speaking distance influences the speech volume first. But it is assumed that the head position difference has the similar effects as the various channels have. So the CMS (Cepstral Means Subtraction)[3] method is used considering its acknowledged validity and acceptable computation complexity for an 8-bit MCU core. Three speakers' speech is tested on the chip set emulation board at real-time, except that one speaker remains the same recognition rate, the other two gain 29% and 55% error rate reduction respectively.

The third factor is the varying background noise, especially other person's speech. We recorded the input phrases through the ADC channel of the chip set emulation board in an office activated with dozens of people and computers. The SNR of the recorded speech is in the range of –2dB~33dB and 78% phrases below 10dB. Under the circumstances, the average recognition rate is on the level of 90%. When the recording is taken in a quiet office, the average SNR is up to above 20dB with 95% phrases above 10dB, and the average recognition rate is up to 96%. Therefore, some noise-robust speech recognition methods should be used in this speech recognition system to broaden its usable condition. But at present, the speech enhancement techniques are so complex for an 8-bit MCU core that the system can only test the SNR of the input phrases and reject the low ones, then give corresponding prompts.

3.3 Segment Mean Template Training and Recognition

The conventional dynamic programming (DP) method, for instance DTW (Dynamic Time Warping) algorithm[2], has been successfully used in the template based speaker-dependent word recognition system and a high recognition rate (over 99%) has been accomplished. But implementing the DTW algorithm in the MCU core of the speech recognition SOC has two obstacles. One is the excessive template storage space. A DTW template of 100 frames, 7 order LPCC each frame, needs 1.4k bytes memory space. This needs too many external data memory space. Another is the template matching process, which it is impossible to carry out in the MCU core real-timely. Therefore the template is not generated directly from each frame of speech as in DTW, the segment mean method of the templates is used. The basic idea is to segment the speech according to the variance of cepstrum and average the feature vectors of each segment. Then concatenating the segment mean feature vectors forms the reference template. Adopting this method, the amount of storage space for templates and template matching space are both reduced to a great extent. The average segment length is 8 frames of speech signal. So about 10 segments are divided for the speech with the duration of 1 second. Then only 280 bytes are needed to store the 7 order LPCC, plus its dynamic variations, of the template. The storage space is only one quarter of those needed previously.

It is known that the template generated from one token only is not robust enough. Besides, it is apt to the endpoint estimation error and segmenting default in this recognition system. So the sequential training method mentioned in [4] is used. The segment mean template is updated in the sequential training. In order to obtain the second consistent token, the user is asked to speak the training phrase second time. Then the first reference template and the second token are compared via a DP process and the resulting distortion score is compared against the threshold to decide whether the second token is consistent with

the reference template. If the score is smaller than the threshold, the second token is accepted to compute the second segment mean template according to the state segmentation obtained by the DP process. Then the final template is computed as the average of two temporary reference templates, and the prompt of successful training is played. If the distortion score is larger than the threshold, the prompt of training failure is played and the phrase can be trained at another time.

By the sequential training method, the performance degradation caused by the segment mean template is compensated quite much. The performance of three different recognition systems that use different templates, i.e., the DTW template, segment mean template and sequential trained segment mean template, is compared on the speech database with 8 mobile phone voice commands of three speakers. The speech data used is recorded through the input real channel of the chip set emulation board, but all the computation is done on the PC platform. The result is given in Table 1.

Table 1. Recognition rate compare of three templates.

Recognition rate	1	2	3
DTW template	100%	99.1%	88.2%
Segment mean template (SMT)	100%	94.4%	82.4%
Sequential trained SMT	100%	97.6%	95.4%

The pronunciation variation of speaker 3 is much more than the other two, so even the DTW template does not work quite well. But the sequential training improves the robustness of template comparatively and the result is satisfying.

The recognition follows two steps. Firstly, distortion scores of the input phrase and the templates are computed via the DP process. Secondly, the class with the minimum measure distance is selected as the recognition result. The Viterbi searching algorithm with path constrained is used, which is considered preferable for the segment mean template matching. In order to remove the redundant computation in the Viterbi searching process, the local score is only computed in a constrained region, as is showed in Figure 4.

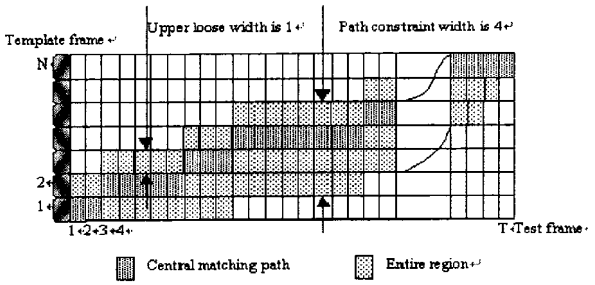

Figure 4. Computation region of path constrained viterbi algorithm.

Two parameters, i.e., upper loose width and path constraint width, control the scope of the searching region. Firstly, the local distortion score of the central matching path, which transits immediately to the next template frame if the score is less than the one computed by the current template frame, with the current test frame is computed. If the distortion score of the current test frame with the current template frame is less, then optimal path still occupies at the current template frame. The upper loose width determines the searching region above the central path and the path constraint width determines the entire region.

It is clear that the global optimal path is included in the constrained region more possibly if the prescribed widths are larger, but the searching process spends more time. The upper loose width has less influence than the path constraint width. When the lower space is unconstrained and the upper loose width equals to 2, there is no segmentation difference compared with the segmentation of the forced alignments of correct words. When the upper loose width equals to 1, there are 3.7% correct tokens having different segmentation compared with the segmentation of the full space Viterbi searching. The percentage of tokens with the error segmentation and the resulting average relative distortion score error under various path constraint widths are illustrated in Figure 5. The actual widths should be the trade-off between the small value and the large value. If the smaller value is used, the DP process can be accomplished in less time, that maybe results in lower recognition rate; If the larger value is used, the more searching paths can be obtained, that results in higher recognition rate.

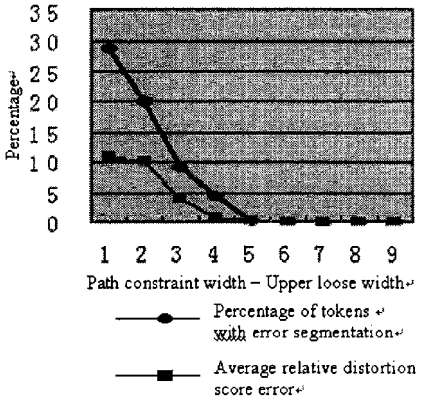

Figure 5. Alignment error and distortion score error under various path constraint widths (upper loose width ≥ 2).

3.4 Speech Coding

Although low-rate parametric coding methods can also produce the high-quality speech, they are too complex for an 8-bit MCU realization. So the simple 16kbits/s or 8kbits/s CVSD is adopted. The prompts are recorded and coded in the PC platform by CVSD, whose parameters are optimized according to the context. Then the data is transferred into the Flash memory. Different from the prompt, the playback must be coded and decoded directly in the chip. The quality of playback is a little less than the prompt.

4. EVALUATION RESULT

The recognition accuracy of the speech recognition SOC is tested on its FPGA emulation board. The testing recognition phrases are short sentences, such as "hello", "what's your name", "how old are you", "do you like some chocolate" and so on. The testing speech database includes 3 speaker speeches. Each speaker speaks 10 different phrases, and each phrase is repeated 30 times. The speech is through the input channel of the FPGA emulation board and is sampled at 8kHz/s. The data collected in this manner just is the real speech input in practical use condition and the recognition accuracy can be assessed exactly. The recognition performances are showed in Table 2 under the different signal to noise ratio conditions.

Table 2. Recognition rates in different SNR conditions.

	Speaker 1	Speaker 2	Speaker 3
0~10dB	90.7%	94.2%	92.7%
10~35dB	95.9%	96.6%	96.6%

It can be seen from Table 2 that the overall recognition accuracy of over 90%, even in the noisy environment. And over 95% speech recognition rate can be obtained when the background SNR is above 10dB. This speech recognition SOC has reached good recognition performance and can finish some simple speech recognition task.

5. SUMMARY

This paper has described the speech recognition SOC with speech recognition and speech coding functions. On the single chip an 8051 compatible MCU core, 512 bytes RAM, 8k bytes ROM, ADC/DAC, PWM, I/O ports and other peripheral circuits are integrated. The characteristics of the chip are summarized in Table 3.

Table 3. Characteristics of the chip.

Process technology	0.5 μm CMOS
Package	64-pin QFP
Operating frequency	40MHz (32768Hz in the idle mode)
Supply voltage	2.4V~5.25V
Power consumption	60mw
Record	27s
Recognition speed	10 phrases with average duration of 1 second in 0.5 second
Recognition accuracy	96.4%

This speech recognition SOC is very cheap and can provide good speech recognition and speech coding performance. This chip can be used for toys, consumer electronics, office devices, and other consuming products. It offers a cheap but modular and flexible solution.

6. ACKNOWLEDGEMENT

This project is supported by National 863 High Technology Projects (Contract No. 863-306ZD13-04-6) of China and National Natural Science Funds (Item No. 69975007) of China. And special thanks go to Li Xiaoyu, who designed and made the chip set emulation board system, those people who developed the FPGA emulation board, and Wen Xue, for offering the CVSD codes.

7. REFERENCES

[1] L.R.Rabiner and R.W.Schafer *Digital processing of speech signals.* Englewood Cliffs, N.J.:Prentice-Hall, 1978.

[2] Xingjun Yang and Huisheng Chi *Speech signal digitla processing.* Publishing House of Electronics Industry, 1995(in Chinese).

[3] Furui S. "Cepstral analysis technique for automatic speaker verification". *IEEE Transactions on Acoustics, Speech, and Signal Processing. roceedings of Eurospeech Conference,* 29(2):254-272, 1981.

[4] Lawrence Rabiner and Biing-hwang Juang *Fundamentals of Speech Recognition.* Prentice-Hall International, Inc., 1993.

BIOGRAPHY

Shi Yuanyuan receives her B.S. degree in electrical engineering in 1997 from Beijing University of Aeronautics and Astronautics. She is currently a Ph.D candidate at Tsinghua University, Beijing, China. Her research focuses upon speech recognition algorithms and ASIC implementation of efficient speech technologies.

Liu Jia receives his B.S., M.S. and Ph.D degrees in communication and electronic systems from Tsinghua University, China, in 1983, 1986 and 1990 respectively. He worked at the Remote Sensing Satellite Ground Station, Academic Sinica, after his Ph.D, and worked as a Royal Society visiting scientist at Cambridge University Engineering Department during 1992-1994. He is now associate professor in the Department of Electronic Engineering, Tsinghua University. His research fields include signal processing, algorithm research, speech recognition, speech synthesis, speech coding and multimedia communication.

Liu Runsheng is a professor in the Department of Electronic Engineering, Tsinghua University. He has engaged himself in the research and tuition on digital circuit, analog circuit, IC design, electronic circuit CAD and speech signal processing.