SPEECH INTERFACE VLSI FOR CAR APPLICATIONS

M. Shozakai

Asahi Chemical Industry Co., Ltd., LSI Labs.
Tanasawa 221, Atsugi, Kanagawa 243-0205, JAPAN
makoto@ljk.atsugi.asahi-kasei.co.jp

ABSTRACT

A user-friendly speech interface for car applications is highly needed for safety reasons. This paper will describe a speech interface VLSI designed for car environments, with speech recognition and speech compression/decompression functions. The chip has a heterogeneous architecture composed of ADC/DAC, DSP, RISC, hard-wired logic and peripheral circuits. The DSP not only executes acoustic analysis and output probability calculation of HMMs for speech recognition, but also does speech compression/decompression. On the other hand, the RISC works as a CPU of the whole chip and Viterbi decoder with an aid of hard-wired logic. An algorithm to recognize a mixed vocabulary of speaker-independent fixed words and speaker-dependent user-defined words in a seamless way is proposed. It is based on acoustic event HMMs which enable a template creation from one sample utterance. The proposed algorithm embedded in the chip is evaluated. Promising results of the algorithm for multiple languages are shown.

1. INTRODUCTION

The concept of ITS(Intelligent Transport Systems) is promoted in many countries. Drivers, cars, roads and information network systems are connected with wireless communication technologies. There must be user-friendly human machine interface for drivers to have an easy access to various information of traffic, road construction, dynamic route guidance for navigation and so forth. The VODIS(Voice Operated Driver's Information Systems) project[1] was launched in Europe to investigate a robust speech interface for command and control applications for car facilities such as a car navigation, a car audio and a cellular phone.

In command and control applications in car environments, some commands("yes", "no", "start" etc.) might be fixed for all users. On the contrary, others(radio station names, voice dialing names etc.) need to be registered as user-defined words. A DTW approach was often used for speaker-dependent speech recognition. A HMM approach is widely used for speaker-independent speech recognition. It is more convenient to use a unified approach for the mixed vocabulary speech recognition. The DTW approach can recognize speaker-independent words and speaker-dependent words in a same fashion. However, a necessity of many templates for speaker-independent words is one disadvantage of processing time for template matching. Another disadvantage is that it is not easy to change the speaker-independent fixed words. We propose the HMM approach to realize the mixed vocabulary recognition without those problems.

This paper is organized as follows. In Section 2 we describe algorithms which are implemented in the proposed chip. In Section 3 a heterogeneous architecture of the chip is described. We propose an algorithm to recognize a mixed vocabulary of speaker-independent fixed words and speaker-dependent user-defined words, based on HMM approach in Section 4. Evaluation results of the algorithm are shown in Section 5. Finally we summarize our proposal and outline our future work.

2. ALGORITHMS

2.1 Speech Recognition

There exist both additive noises and multiplicative distortions in car cabin environments. The additive noises are classified into two categories. One is a known additive noise source signal of that is available like car audio speaker output. The other is an unknown additive noise source signal of that is not available like noises generated by engine, road friction, wind and air conditioner. The multiplicative distortions include microphone characteristics, acoustic transmission characteristics from mouth to remote microphone, speaking style and speaker personality that depends on vocal tract and vocal cords. We proposed three algorithms to cope with these environmental noises.

(1)NLMS-VAD(Normalized Least Mean Squares with frame-wised Voice Activity Detection)[2] is able to cancel effectively the known additive noise which is embedded in the unknown additive noise. A precise control of FIR filter coefficients update with frame-wise VAD and a save/restore mechanism of those coefficients enables higher ERLE(Echo Return Loss Enhancement) than a conventional NLMS with sample-wise DTD(Double Talk Detection).

(2)CSS(Continuous Spectral Subtraction)[3] can suppress stationary unknown additive noises. It is also capable of masking weak residual of known additive noise that can not be sufficiently cancelled by NLMS-VAD.

(3)E-CMN(Exact Cepstrum Mean Normalization)[4] is able to compensate multiplicative distortions simultaneously. It works as a dynamic frequency equalizer to extract normalized spectra.

An acoustic analysis in the chip is processed as follows. A sampling frequency is 12kHz. At first the known additive in microphone input is cancelled by NLMS-VAD with 190 taps FIR filter if its source signal, which is a mixed signal of left and right channels of car audio, is available. Then, the stationary unknown additive noise is suppressed with CSS. 10 MFCC(Mel Frequency Cepstrum Coefficients)s, 10 delta MFCCs and 1 delta log power is extracted every 10ms frame shift with 21.3ms frame length. The multiplicative distortions are compensated by E-CMN applied to the MFCCs.

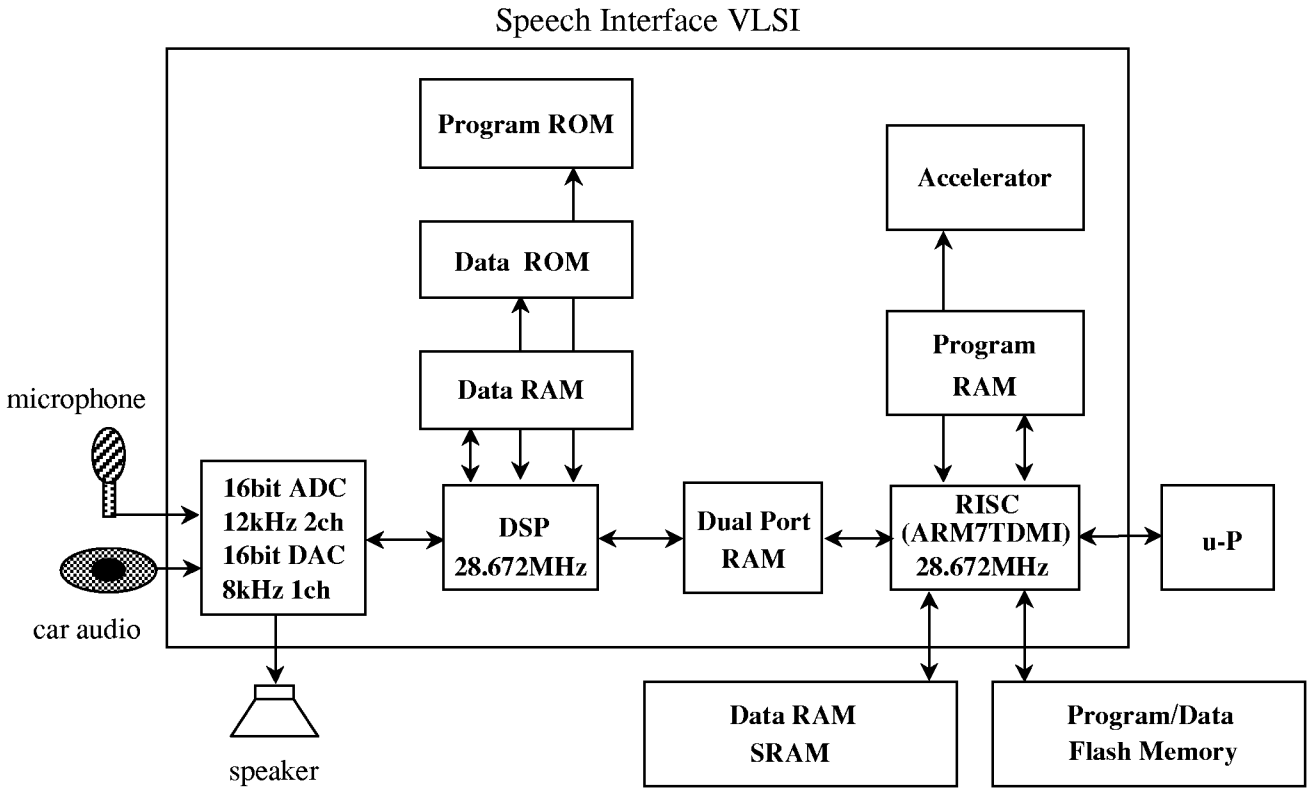

Fig.1: Block diagram of the chip.

The chip uses 54 Japanese phoneme models of tied-mixture HMM(TM-HMM). Each phoneme model has three emission states as shown in Fig.2(a). Numbers of Gaussian distributions shared by all states in all phoneme HMMs are 256, 256 and 64 to MFCCs, delta MFCCs and a delta log power respectively.

Word matching is done by Viterbi decoding algorithm. A beam search technique is used to make a response time after an end of utterance within 1 second.

2.2 Speech Compression/Decompression

A speech compression/decompression algorithm called MMEV(Multimodal Multipulse Excitation Vocoder)[5] is equipped in the chip. It is a variable rate speech codec with selectable bit rate from 4.5kbps to 13kbps. The MMEV is used to generate encoded talk-back speech data for speaker-dependent user-defined words and to play system guidance and talk-back speech data.

3. ARCHITECTURE

The proposed speech interface VLSI is realized as a heterogeneous architecture which is composed of 16bits ADC/DAC, a proprietary 16bits DSP, a licensed 32bits RISC(ARM7TDMI), hardware logic and peripheral circuits. A block diagram is shown in Fig.1. The chip works as a slave processor to a host microprocessor at 28.672MHz at 5V.

(1)The ADC has 2 channels of 16 bits, one for microphone, the other for source signal for adaptive filtering for car audio speaker output. A SNR of ADC is about 60dB. The DAC has 1 channel of 16 bits which is used for playing system guidance and talk-back speech data.

(2)The DSP with dual MACs(Multiply and ACcumulations) executes signal processing as follows. In modes of word recognition and word registration, the acoustic analysis described in Subsection 2.1 and log probability calculation of shared Gaussian distributions in TM-HMMs are executed. It is to be noted that it can execute one update of FIR filter coefficient by NLMS with two machine cycles. Furthermore, it executes MMEV compression in mode of generating talk-back speech and MMEV decompression in mode of playing system guidance and talk-back speech, too. All of data and program ROM and data RAM for the DSP are embedded in the chip.

(3)The RISC operates as not only a chip controller but also a output probability calculator and Viterbi decoder for the TM-HMMs. An external FLASH memory stores programs of RISC, parameters of phoneme TM-HMMs, templatess of fixed-words and user-defined words, encoded data of system guidance and talk-back speech of user-defined words etc. An external SRAM is used as a working area for template creation of user-defined words and for word matching by Viterbi decoding. An internal RAM is used as an overlay area to which a segment of RISC program is repeatedly downloaded to realize no-wait program fetch.

(4)An ACS(Add-Compare-Select) processor and a memory addressing unit for fast access of weight parameters of the TM-HMMs are added as a hardwired logic to accelerate Viterbi decoding.

(5)The peripheral circuits not shown in Fig.1 include MMU(Memory Management Unit), PLL(Phase Lock Loop) and etc..

4. MIXED VOCABURARY RECOGNITION

It is necessary to create a template of speaker-dependent user-defined word to realize mixed vocabulary recognition. From a user's point of view, it is desirable to be able to create the template with only one utterance. A technique to train Gaussian distribution's mean vectors of user-defined word HMM with one covariance matrix shared is widely applied[6]. It is sometimes difficult to train the word HMM with just one sample utterance in a statistical sense. An alternative is to activate a phoneme recognizer[7]. The obtained phoneme sequence can be interpreted as an approximation of the sample utterance and used as the template. But, a constraint of three successive emission states in each phoneme HMM seems to prevent more precise approximation.

All phonemes have multiple acoustic events. It seems reasonable to interpret that each state in phoneme HMM represents one of acoustic events. Three HMMs which are split from phoneme HMM as shown in Fig.2(b) are called acoustic event HMMs. We propose the acoustic event recognizer is activated to create the template of user-defined word instead of the phoneme recognizer.

A network in Fig.3(a) accepts a sequence of any length and any order of acoustic events, where "X.Y" means that X is a phoneme label and Y is a state number. However, it might consume huge amount of Viterbi decoding time by combinatorial explosion even if a beam search is applied. So, we use a simplified network of acoustic events in Fig.3(b) to avoid the problem. We assume it is not frequent that acoustic event HMMs having state number '1' resides at state number '3' position. This simplification enables much reducing of number of acoustic event combination.

Obtained acoustic event sequences for speaker-dependent user-defined words can be easily added to a vocabulary network of phoneme sequence for speaker-independent fixed words.

Viterbi decoding with the extended vocabulary network is executed to recognize the mixed vocabulary.

5. EVALUATION

5.1 Multilingual Speaker Dependent Recognition

The proposed chip is evaluated to a task of speaker dependent word recognition for Japanese, English, French, German and Italian with using Japanese phoneme HMMs. The vocabulary size is 128 words. Two sample utterances per each word with car environmental noise are collected. One sample of SNR 20dB is used to create word template as proposed in Section 4. The other of SNR 10dB is used as evaluation data. Table 1 shows number

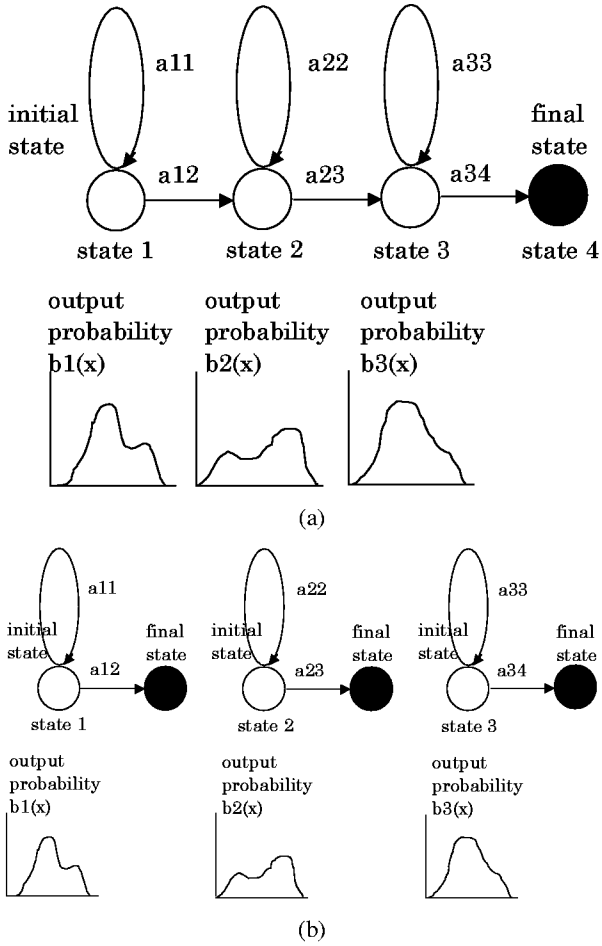

Fig.2: Split into acoustic event HMMs.

of speakers, average recognition rate and lowest recognition rate. It is observed that rather high recognition performance can be obtained for every language. It is interesting to see that German and Italian get higher performance than English and French. It reflects similarities of phoneme structure between Japanese and other languages.

Table 1: Performance of multilingual speaker dependent word recognition with Japanese phoneme HMMs.

	no. of speakers	average	lowest
Japanese	6(2 females/4 males)	97.3%	95.3%
English	9(3 females/6 males)	92.2%	87.5%
French	5(2 females/3 males)	93.4%	89.8%
German	6(2 females/4 males)	95.8%	93.8%
Italian	4(2 females/2 males)	95.1%	93.8%

5.2 Japanese Mixed Vocabulary Recognition

Second evaluation of the chip is a task of mixed vocabulary word recognition for Japanese. The vocabulary size is 128 words. Phoneme sequences of 41 fixed words are compiled in advance in vocabulary network. One sample utterance with car noise of SNR 20dB per 87 user-defined words is used to create a template that is added to the vocabulary network. One sample utterance

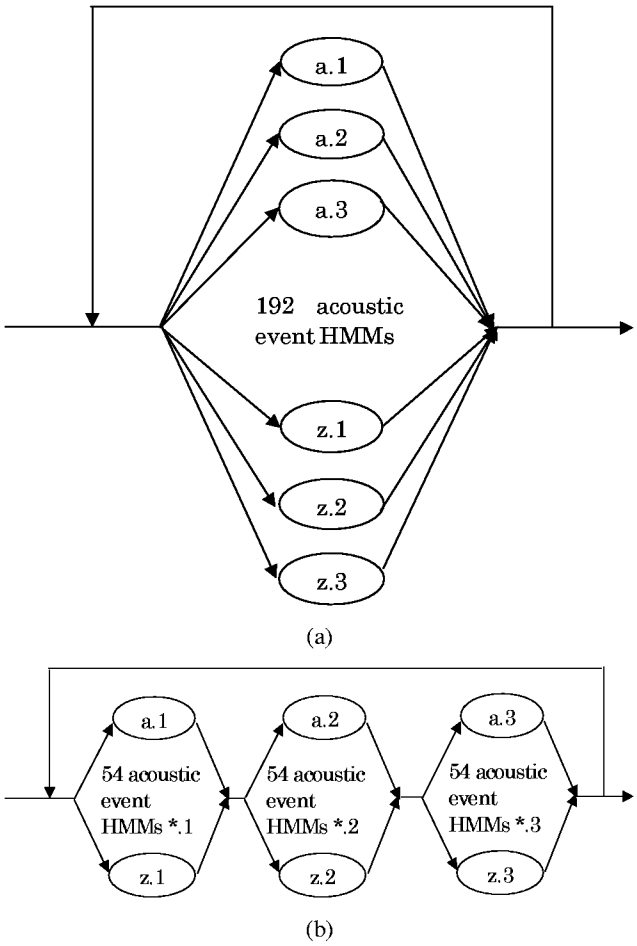

Fig.3: Syntax of acoustic event HMMs.

with car noise of SNR 10dB per each word is used as evaluation data. Table 2 shows number of speakers, average recognition rate and lowest recognition rate. A combination of phoneme sequence for fixed word and acoustic event sequence for user-defined word works well for Japanese mixed word vocabulary.

Table 2: Performance of Japanese mixed vocabulary word recognition.

	no. of speakers	average	lowest
Japanese	6(2 females/4 males)	92.2%	89.8%

5.3 Multilingual Mixed Vocabulary Recognition

The last evaluation is a task of multilingual mixed vocabulary word recognition by the proposed algorithm for English which gives lowest average performance and German which gives highest average performance except Japanese in Table 1. It is generally hard to invent phoneme sequence of fixed word for English and German with knowledge of Japanese phonemes. Instead, acoustic event HMM sequences which are extracted from 4 speakers(2 females/2 males) sample utterances of SNR 20dB for 16 fixed words are in advance registered in the vocabulary network by the analogy with DTW-based multi templates approach. One sample utterance of SNR 20dB per 64 user-defined words is used to create a template of user-defined words. The vocabulary size is totally 80 words. One sample utterance per 80 words with car noise of SNR 10dB is used as an evaluation data. Speakers whose utterances are used to create templates of fixed words are excluded from the evaluation data. Table 3 shows number of speakers, average recognition rate and lowest recognition rate. Multi template approach for fixed words seems promising. The template of acoustic event HMM seems richer and more compact representation of inter-speaker variation than a template of acoustic parameter in DTW-based recognizer.

Table 3: Performance of multilingual mixed vocabulary word recognition with Japanese phoneme HMMs.

	no. of speakers	average	lowest
English	5(1 female/4males)	94.0%	87.5%
German	2(0 female/2 males)	96.3%	95.0%

6. SUMMARY

This paper described the speech interface VLSI of a heterogeneous architecture for car applications. It has functions of mixed vocabulary word recognition and a variable rate speech compression/decompression. The algorithm of mixed vocabulary word recognition was proposed. It was shown that construction of multilingual speaker-dependent word recognizer and Japanese mixed vocabulary word recognizer with a set of Japanese acoustic event HMMs is feasible. It was also shown that it is promising to realize multilingual mixed vocabulary word recognizer with a set of Japanese acoustic event HMMs with a multi template technique. Next research goal is to investigate how to design a universal set of acoustic event HMMs for multilingual mixed vocabulary speech recognition.

7. REFERENCES

[1] Pouteau, X., Krahmer, E. and Landsbergen, J., "Robust Spoken Dialogue Management for Driver Information Systems", *Proc. EUROSPEECH, pp.2207-2210,* Rhodes, Greece, 1997.

[2] Shozakai, M., Nakamura, S. and Shikano, K., "Robust Speech Recognition in Car Environments", *Proc. ICASSP, pp.269-272,* Seattle, 1998.

[3] Shozakai, M., Nakamura, S. and Shikano, K., "A Speech Enhancement Approach E-CMN/CSS for Speech Recognition in Car Environments", *Proc. IEEE Workshop of ASRU,* Santa Barbara, 1997.

[4] Shozakai, M., Nakamura, S. and Shikano, K., "A Non-iterative Model-Adaptive E-CMN/PMC Approach for Speech Recognition in Car Environments", *Proc. EUROSPEECH, pp.287-290,* Rhodes, Greece, 1997.

[5] Unno, T., Barnwell III, T., P., Clements, M. A., "The Multimodal Multipulse Excitation Vocoder", *Proc. ICASSP, pp.1683-1686,* Munich, 1997.

[6] Viikki, O., Bye, D. and Laurila, K., "A Recursive Feature Vector Normalization Approach for Robust Speech Recognition in Noise", *Proc. ICASSP, pp.733-736,* Seattle, 1998.

[7] Ramabhadran, B., Bahl, L. R., deSouza, P. V. and Padmanabhan, M., "Acoustics-only Based Automatic Phonetic Baseform Generation", *Proc. ICASSP, pp.309-312,* Seattle, 1998.