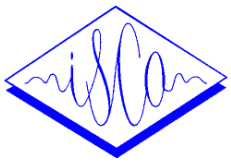

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A Keyword-Spotting Unit for Speaker-Independent Spontaneous Speech Recognition

Yasuyuki MASAI†, Jun'ichi IWASAKI†, Shin'ichi TANAKA†, Tsuneo NITTA†
Masahiro YAO‡, Tomohiro ONOGI‡, Akira NAKAYAMA‡

†Multimedia Engineering Laboratory, TOSHIBA Corporation, Japan
‡TOSHIBA Computer Engineering Corporation, Japan

ABSTRACT

In this paper, we describe a real-time keyword-spotting unit (KeySpot) with an adaptive noise-canceller for speaker-independent, spontaneous speech recognition in noisy environments.

KeySpot consists of a DSP (TMS320C30) for adaptive noise-cancellation and acoustic analysis, a special LSI for statistical matrix quantization (SMQ), two SPARC chips ('SPARC1' and 'SPARC2') for HMM based keyword-spotting, and a SPARC chip ('SPARC3') for syntactic analysis.

KeySpot was tested under two conditions: a speaker-independent large-vocabulary isolated word recognizer, and a speaker-independent small-vocabulary word spotter. Evaluation results have shown that KeySpot can be used for the speaker-independent, 1000 isolated word recognizer with an accuracy of 96.3%, as well as the 90 word vocabulary word spotter with an accuracy of 94.4% with a response time of 0.3 sec.

1 INTRODUCTION

At present, social-automation systems, such as a directory guidance system, a ticket-vending machine, or an ATM are widely used. However, because there are limitations in user interface of these systems, a human-oriented man-machine interface is expected. The authors have developed a multimodal dialogue system[1] which provides multiple input channels of spontaneous speech and touch, as well as multiple output channels of graphics and voice response.

Because the social-automation system is used by unspecific users, the speech recognition in the system should deal with the problem of speaker variability and spontaneous speech, which includes nonverbal noise such as tongue clicking, insertion of irrelevant words, self-correction, abbreviated expressions, etc. Moreover, severe environment requires a speech recognition unit to provide robustness against noise. The response time to the input of speech is another important item for the user interface of the system.

In this paper, we describe a newly developed keyword-spotting unit ("KeySpot"). KeySpot can recognize 500 keywords in continuous speech or 1000 isolatedly spoken words. KeySpot is based on SMQ/HMM (a Statistical Matrix Quantization and discrete HMM hybrid algorithm), which incorporates pattern variations of a phonetic segment level and a word level and performs keyword-spotting to handle sponteneous speech[2]. KeySpot is also capable of cancelling noise by using adaptive noise cancellation techniques with two microphones[3]. The performance of the noise reduction is typically 10dB. Section 2 provides an SMQ/HMM hybrid algorithm and section 3 explains KeySpot architecture. Finally, section 4 presents experimental results for isolated word recognition and keyword-spotting.

2 OVERVIEW OF AN SMQ/HMM HYBRID ALGORITHM

The recognition system based on SMQ/HMM[4] is divided into two large parts: the phonetic segment stage, which performs statistical matrix quantization (SMQ), and the word stage, which uses the improved HMM training algorithm. We expect to incorporate fine phonetic variations less than 100 msec. into an orthogonalized phonetic segment codebook of the SMQ, as well as speech variations more than 100 msec. into word HMMs.

2.1 Phonetic Segment

Various types of speech events are observed in continuous speech. Some types can be described only by a VCV unit and other types by an acoustic segment. Therefore, we need to use multiple phonological units for speech description. A phonetic segment[4] extracted from a Japanese speech database consists of about 700 acoustic/phonetic structures with duration varying between 32 and 96 msec. (e.g. acoustic segment, phoneme (C, V), C_V, CC, $_{VC}$ *and* $_VC_V$).

2.2 SMQ and an Orthogonalized Phonetic Segment Codebook

The SMQ effectively incorporates pattern variations of each phonetic segment into an orthogonalized codebook, or an eigen vector set, using the Karhunen Loeve Transform. The matching score or similarity S_{ic} between the orthogonalized codebook V_{rc} of phonetic segment c and a normalized input pattern $X_i = (x_{11}, \ldots, x_{nt}, \ldots, x_{NT})$ is defined as follows:

$$S_{ic} = \sum_{r=1}^{R} W_r (X_i \cdot V_{rc})^2 \qquad (1)$$

where W_r are weight coefficients, $(\cdot)$ denotes inner product and $R(= 8)$ is the number of eigen vectors. Equa-

tion 1 is the same expression used in the Multiple Similarity Method[5] and the Sub-space Method[6].

2.3 SMQ/HMM

The hidden Markov model investigated in this paper is a left-to-right model of a discrete density HMM. Transition probabilities and output probabilities are trained with K-best codes in SMQ[7] and estimated by using the forward-backward algorithm[8]. An optimal state sequence in HMM networks is searched with the Viterbi algorithm. The SMQ/HMM has achieved a high performance on a speaker-independent and large-sized vocabulary word recognition tasks[7].

2.4 Interleaf Model

We proposed a new sub-word HMM structure style (Interleaf model)[9] for continuous, spontaneous, keyword-based, speaker-independent speech recognition. In the Interleaf model, two types of phones appear alternately in a word. The first type of phone is a monophone having a base-only part, and the second type of phone is a diphone having a transition-only part. All models are initialized with a uniform distribution. No initial segmentation is performed at all. All training models are concatenated from shared, untrained subword models. Segmentation is performed automatically by the forward-backward algorithm during normal training. The Interleaf model was an improvement over a word-based model on a 227 closed-words test and a 22 open-words test[9].

3 A KEYWORD-SPOTTING UNIT (KeySpot)

A speaker-independent, keyword-spotting unit (KeySpot) was designed for recognizing 500 keywords in continuous speech or 1000 isolated spoken words in real time. The response time of 0.3sec. is for an end point detection.

Figure 1 shows a block diagram of KeySpot based on SMQ/HMM. Table 1 shows a specification of KeySpot. KeySpot consists of a DSP(TI: TMS320C30, 33MHz) for adaptive noise-cancellation and acoustic analysis, a special LSI (Phonetic Segment Engine Chip : PSE) for SMQ, two SPARC chips (Fujitsu: SPARClite, 33MHz 'SPARC1' and 'SPARC2') for HMM based keyword-spotting, and a SPARC chip ('SPARC3') for syntactic analysis.

3.1 Adaptive Noise-Cancellation and Acoustic Analysis Module

Adaptive noise-cancellation (ANC) using two microphones and acoustic analysis of 16 LPC mel-cepstrum every 8 msec. are executed by a DSP. The DSP has 128Kbytes of SRAM and is connected to a syntactic analysis module with a FIFO memory. Figure 2 shows a block diagram of ANC using the Least-Mean Square (LMS) adaptation algorithm. In a speech recognition application, control of ANC renewal is important because speech is inputted both on a main microphone and a reference microphone. ANC renewal, or starting/stopping the renewal of coefficients of the adaptive filter, is controlled by discriminating between voice and ambient noise according to the power of an input speech s_i from the main microphone and noise n_i from the reference microphone, and a residual signal e_i[10]. The control of ANC renewal stops adaptation while a voice is detected. The residual signal e_i is used for an input in the stage of LPC analysis. Figure 3 shows an example of a contaminated speech signal s_i, a residual signal e_i and an index of noise reduction ($10\log(\overline{e_i^2} \,/\, \overline{s_i^2})$). The example was tested under the noisy environment of 80dBA and 64 taps FIR filter was used for ANC.

Table 1: Specification of KeySpot.

Recognition Algorithm	SMQ/HMM
Speaker	Speaker-Independent
Speaking Sytle and Vocabulary	Isolated Word: 1000 words Continuous Speech: 500 words
Sampling Characteristics	12kHz, 16bits
Response Time	less than 0.3 sec.
Interface	RS-232C

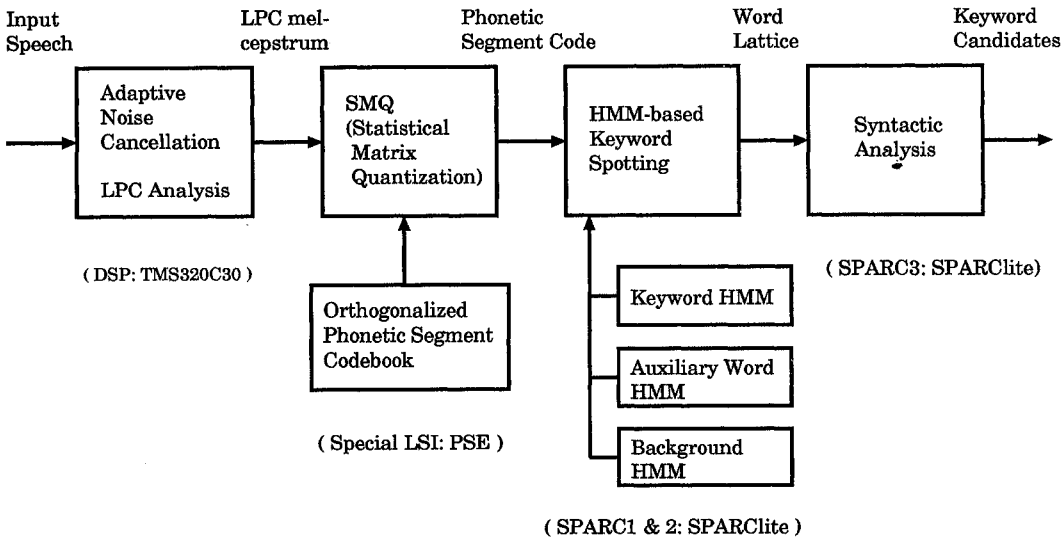

Figure 1: Block Diagram of a Keyword-Spotting Unit (KeySpot)

3.2 Statistical Matrix Quantization Module

A special LSI (PSE) for SMQ rapidly converts speech parameters (16 LPC mel-cepstrum sequences) into phonetic segment codes by using the equation (1). Table 2 shows the specification of the PSE. The PSE executes 4.8×10^8 multiply/accumulate instructions per second. The phonetic segment codebook is stored on 560 Kbyte of DRAM. The PSE is connected directly to the DSP with an on-chip parallel I/O port.

3.3 Hidden Markov Model Module

A HMM module which is composed of two SPARC chips searches an optimal state sequence in word HMMs with the Viterbi algorithm. We use two normalization procedure to normalize keyword likelihood score, namely, likelihood score normalization using a background HMM[11] and duration normalization. The structure of the background HMM is the same as that of the word model. Each SPARC chip has 16Mbytes of DRAM and 2Mbytes of EPROM, and is connected to a syntactic analysis module with a two-way FIFO memory.

3.4 Syntactic Analysis Module

A syntactic analysis module consists of a SPARC chip, 16Mbytes of DRAM, 2Mbytes of EPROM and two serial I/O ports. The module is also used for a main controller of KeySpot and excutes end-point detection, keyword lattice parsing to find a best word sequence[12] and One-Pass search using finite state networks[13]. Because the syntactic analysis module can execute the Viterbi algorithm for a small-sized vocabulary word by itself, KeySpot works without the HMM module.

4 EXPERIMENTS

4.1 Speech Database

Four datasets were used for training and the evaluation of isolated word recognition and keyword-spotting.

The first dataset $DS-1$ was used for designing codebook in SMQ. $DS-1$ includes 250 phonetically balanced words uttered by 15 male and 15 female speakers. 40,500 segments were manually extracted to design a codebook which includes 652 full Japanese phonetic segments.

The second dataset $DS-2$ was used for HMM training. $DS-2$ was composed of 1000 isolated words (station name, digit, place name, etc.) uttered by 25 male and 25 female speakers.

The third dataset $DS-3$ was used for isolated word recognition testing. $DS-3$ was a 773-word subset of $DS-2$ uttered by unknown speakers (5 males and 5 females).

The forth dataset $DS-4$ was used for keyword-spotting testing in continuous speech. $DS-4$ was composed of 40 sentences uttered by unknown speakers (5 males and 5 females). The sentences contained extraneous speech "eeto (Um)" before a keyword and "onegaishimasu (please)" after the keyword.

Table 2: Specification of Phonetic Segment Engin Chip (PSE)

Number of Gate	300KGates
Input Pattern	Mel-Cepstrum Sequences Data Length: 8bits Maximum Dimension: 512
Reference Pattern	Phonetic Segment Codebook Data Length: 8bits Maximum Dimension: 512 Maximum Number of Eigen Vectors: 32
Performance	Multiply/Accumulate: 50 nsec. 24 Parallel Multipliers
Output	Sorted Phonetic Segment Codes and Scores within the Top 15 Data Length: 16bits

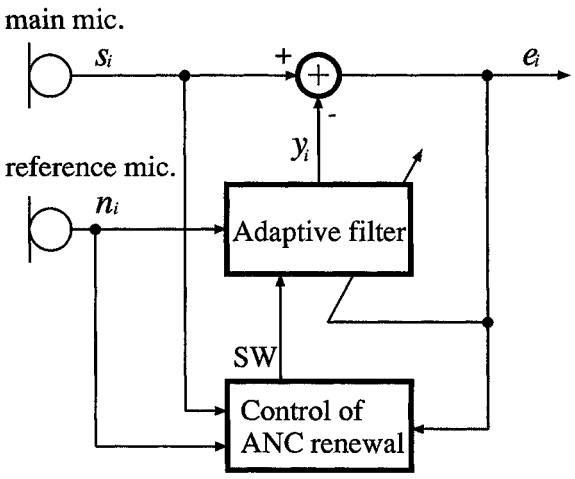

Figure 2: Block Diagram of an Adaptive Noise-Canceller

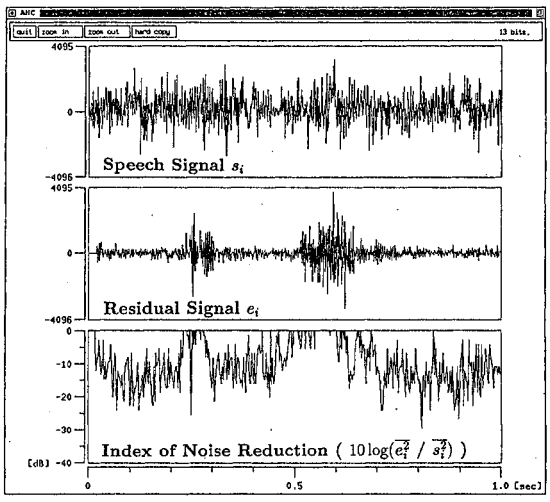

Figure 3: An Example of Noise Reduction with ANC

4.2 Isolated Word Recognition and Keyword-Spotting

We performed tests on 1000 isolated word recognition using the dataset $DS-3$ and achieved recognition accuracy of 96.3%. HMMs used in all the experiments of this section were word models with 10 states.

In the keyword-spotting task with 90 vocabulary, we examined the following two keyword-spotting algorithms using the dataset $DS-4$:
(1) One-Pass search with finite state networks(cf. figure 4)
(2) keyword lattice parsing.

Three filler models were tested in the One-Pass search algorithm. The first filler model (Filler1) uses a garbage model which is made from 546 words out of the vocabulary. The second filler model (Filler2) has two word-models of "eeto (Um)" and "onegaishimasu (please)" in addtion to the garbage model of Filler1. The third filler model (Filler3) has only two word-models of "eeto (Um)" and "onegaishimasu (please)".

The experimental results are shown in table 3. Filler3 achieved the best recognition accuracy of 94.4%, however, because various nonvocabulary items will occur in practical applications, frequently occuring nonvocabulary items should be properly incorporated into filler models, as well as the garbage model for nonfrequent nonvocabulary items. In the keyword lattice parsing algorithm, false-alarm error was the main factor of mis-recognition.

5 CONCLUSION

A real-time keyword-spotting unit (KeySpot) with an adaptive noise-canceller was presented. KeySpot can be used for the 1000 isolated word recognizer as well as the 500 word spotter.

In future work, sub-word HMMs will be evaluated for a vocabulary unspecific word-spotter. We also have a plan to evaluate KeySpot in noisy environments, and to apply it to social-automation systems in the real world.

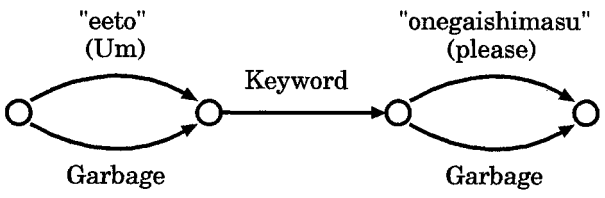

Figure 4: Finite State Network for One-Pass Search

Tabel 3: Recognition Accuracies for Isolated Word Recognition and Keyword-Spotting in Continuous Speech.

Spotting Algorithm	Isolated Word (90 words)	Continuous Speech (90 words)
One-Pass Search with Finite State Networks		
Filler1	93.5%	79.2%
Filler2	93.0%	91.9%
Filler3	96.0%	94.4%
Keyword Lattice Parsing	92.2%	75.6%

References

[1] T.Nitta, Y.Masai, J.Iwasaki, S.Tanaka, H.Kamio and H.Matsu'ura, "A Multimodal Directory Guidance System with an Interractive Mechanism", Proc. EUROSPEECH 93, pp.2055-2058, 1993.

[2] Y.Masai, S.Tanaka and T.Nitta, "Speaker-Independent Keyword Recognition Based on SMQ/HMM", Proc. ICSLP92, pp.619-622, 1992.

[3] T.Nitta, S.Minami, A.Nakayama and T.Onogi, "Speech Recognition under Noisy Environment", Trans. Speech Research, IEICE Tech. Report, SP94-20, pp.45-52, 1994 (in Japanese).

[4] T.Nitta, J.Iwasaki and H.Matsu'ura, "Speaker Independent Word Recognition using HMMs with an Orthogonalized phonetic Segment Codebook", Proc. EUROSPEECH 91, pp.1107-1110, 1991.

[5] T.Nitta, T.Murata, H.Tsuboi, T.Kawada and S.Watanabe, "Development of Japanese Voice-activated Word Processor using Isolated Monosyllable Recognition", Proc. ICASSP82, pp871-874, 1982.

[6] E.Oja, "Subspace Method of Pattern Recognition", Research Studies Press, 1983.

[7] T.Nitta, J.Iwasaki, Y.Masai and H.Matsu'ura, "Representing Dynamic Features of Phonetic Segment in an Orthogonalized Codebook of HMM Based Speech Recognition System", Proc. ICASSP92, pp.385-388, 1992.

[8] L.R.Bahl, F.Jelinek and R.Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.PAMI-5, No.2, pp.179-190, 1983.

[9] M.Pundsack, T.Nitta, "Comparison of Context Dependent Sub-word HMMs for Japanese", Trans. Speech Research, IEICE Tech. Report, SP93-112, pp.63-70, 1993.

[10] S.Minami and T.Kawasaki, "A Double Talk Detection Method for an Echo Canceller", IEEE ICC'85, pp.1492-1497, 1985.

[11] R.C.Rose and D.B.Paul, "A Hidden Markov Model Based Keyword Recognition System", Proc. ICASSP90, pp.129-132, 1990.

[12] Y.Takebayashi, H.Tsuboi, Y.Sadamoto, H.Hashimoto and H.Shinchi, "A Real Time Speech Dialogue System using Spontaneous Speech Understanding", Proc. ICSLP92, pp.651-654, 1992.

[13] A.Kai and S.Nakagawa, "A Frame-Synchronous Continuous Speech Recognition Algorithm using a Top-Down Parsing of Context-Free Grammar. Proc. ICSLP92, pp.257-260, 1992.