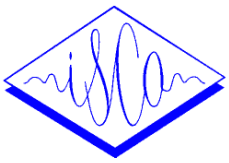

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Speaker–Independent Keyword Recognition Based on SMQ/HMM

Yasuyuki Masai, Shin'ichi Tanaka and Tsuneo Nitta

Information Systems Engineering Laboratory, TOSHIBA Corp.
70 Yanagi–cho, Saiwai–ku, Kawasaki, 210, JAPAN

ABSTRACT

This paper describes a speaker-independent keyword recognition system based on hidden Markov models (HMM's). We propose a new matrix quantization (MQ) algorithm called Statistical MQ (SMQ) that uses an orthogonalized phonetic segment codebook and a word beginning frame prediction (BFP) algorithm to achieve accurate and efficient word-spotting.

The SMQ effectively incorporates pattern variations of each phonetic segment into the orthogonalized phonetic segment codebook containing about 700 phonetic segments, and transforms the input speech into a sequence of phonetic symbols. The BFP algorithm predicts a word beginning frame in which the next Viterbi alignment should be generated, using the transition frame at which the initial transition to the second state occurred in the most recent Viterbi alignment.

The proposed keyword recognition system has been tested on a data set of 32 keywords, 5 auxiliary words, and 17 unknown words. The test data is compound words uttered by 6 unknown speakers and the keyword recognition accuracy was 91.1%.

10.21437/ICSLP.1992-207

1 INTRODUCTION

We have previously developed a speaker-independent, small vocabulary word recognition system which can be used in noisy environments for such applications as a voice-activated ticket vending machine and a voice-input elevator[1,2]. However, such small vocabulary systems cannot compete with conventional key-input systems for applications other than those requiring hands-free, attention-free operation, and for which voice activation is relatively easy to use. Large vocabulary systems, on the other hand, can benefit most from the advantages of "direct access" voice activation, and we previously announced the development of such a real-time recognition system[3].

In speaker-independent systems used by many people, such as in information services, it is essential that they tolerate irrelevant words, self-correction, abbreviated expressions, and other traits of spontaneous speech. In particular, because compound proper nouns such as place names and building names are abundant, and since formal usage is rare, abbreviated speech is quite common. To deal with this, we introduce a word-spotting approach in which compound words are divided into keywords and auxiliary words, and the words are recognized by HMM-based word spotting[4].

The following three items are vital for continuous word spotting:
1) Spotting the correct end-points of words;
2) Recognizing keywords accurately;
3) Reducing the computing time.
A large margin in end-point detection increases the number of possible word sequences. In addition, a huge number of calculations are needed for continuous word spotting. To cope with these problems, we propose a BFP algorithm to reduce computation time without decreasing recognition accuracy, and report that we have achieved accurate keyword spotting through the use of both likelihood-score normalization using a background HMM[5], and syntactic analysis of an input word lattice consisting of keywords and auxiliary words[6].

Section 2 gives a description of the SMQ/HMM, section 3 explains the keyword-spotting algorithm, and section 4 examines our proposed method.

2 OVERVIEW OF THE SMQ/HMM HYBRID ALGORITHM

A block diagram of the recognition system based on the SMQ/HMM[4] is shown in Fig. 1. The system operates on the phonetic segment level, which performs statistical matrix quantization (SMQ), and on the word level, which uses an improved HMM training algorithm. We expect to incorporate fine phonetic variations of 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 SEGMENTS

Many types of speech events are observed in continuous speech. Some can only be described by a VCV unit and others types by an acoustic segment. Therefore, we need to use multiple phonological units for speech description. The phonetic segment[4] extracted from a database of Japanese speech consists of about 700 acoustic/phonetic structures with durations varying between 32 and 96 msec. (e.g. acoustic segment, phoneme(C,V), Cv, CC, vc, vCv).

2.2 SMQ AND ORTHOGONALIZED PHONETIC SEGMENT CODEBOOK

The SMQ effectively incorporates pattern variations of each phonetic segment into an or-

thogonalized codebook, or an eigenvector set, using the Karhunen-Loeve Transform. The matching score, or similarity S_{ic} between the orthogonalized codebook V_{rc} of a phonetic segment c and a normalized input pattern $X_i = (x_{11},\cdots,x_{nt},\cdots,x_{NT})$ is defined as follows :

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

where $(\cdot)$ denotes the inner product and R (=8) is the number of eigenvectors. Equation (1) is the same expression as used in the Multiple Similarity Method [7] and the Sub-space Method [8].

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[3] and estimated using the forward-backward algorithm[9]. The optimal state sequence in the HMM network is searched with a Viterbi algorithm. The SMQ/HMM has high performance in a speaker-independent, large vocabulary, and isolated word recognition task[3].

3 KEYWORD SPOTTING

3.1 BEGINNING FRAME PREDICTION (BFP) ALGORITHM

In most word-spotting systems, a Viterbi alignment is generated for every pair of beginning and ending frames; however, because a large number of word candidates are produced, the resultant keyword recognition accuracy is decreased.

In our system, the BFP algorithm effectively predicts the beginning frame FS(t) in which the next Viterbi alignment should be generated. Fig. 2 shows the word-spotting scheme using the BFP algorithm. First, we obtain the transition frame FØ1(t-1) at which the initial transition to the second state occurred in the most recent Viterbi alignment, by back-tracing the optimum Viterbi path. Next, we predict a beginning frame FS(t) using the beginning frame FS(t-1) and the transition frame FØ1(t-1) in the most recent Viterbi alignment. We define the prediction rule as follows:

$$FS(t)=\begin{cases} FS(t-1)+J\emptyset1(k) & (F\emptyset1(t-1)-FS(t-1) = \emptyset) \\ F\emptyset1(t-1) & (\emptyset < F\emptyset1(t-1)-FS(t-1) \leqq J\emptyset1(k)) \\ F\emptyset1(t-1)-J\emptyset1(k) & (F\emptyset1(t-1)-FS(t-1) > J\emptyset1(k)) \end{cases} \qquad (2)$$

where t is the number of Viterbi alignment calculations , k is the category number, and JØ1(k) is the average frame length in the initial state for category k. The ending frame that has the best likelihood is given by calculating the Viterbi alignment.

3.2 BACKGROUND HMM

The keyword likelihood score in continuous speech often fluctuates considerably. We use a background likelihood score in addition to duration normalization to normalize the keyword likelihood scores[5]. The structure of the background HMM is the same as that of the word model. First, background HMM training is performed for each word. Then, averaged transition probabilities and output probabilities are used as background HMM parameters. A normalized log-likelihood score of a word (S_n) is obtained as follows:

$$S_n = \alpha S_k - (\alpha - 1) S_b \qquad (3)$$

where S_k is the log-likelihood score of the word HMM and S_b is the log-likelihood score for the overlapping string of the background HMM. Because both S_k and S_b tend to have lower scores at transients in the input speech, S_n in equation (3) is expected to give a good normalized score.

3.3 RECOGNITION OF A KEYWORD IN A COMPOUND WORD

Compound proper nouns are often pronounced in various abbreviated forms. To deal with this problem, we propose a keyword recognition method which can recognize a keyword embedded in a compound word. The following is our approach.

First, all the compound words are divided into keywords and auxiliary words. For example, in "TOKYO TOSHIBA BUILDING" ,"TOSHIBA" is a keyword,

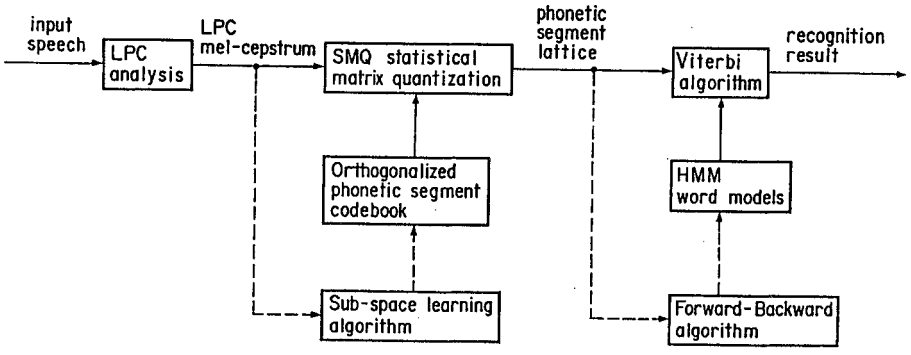

Figure 1. A block diagram of the recognition system based on SMQ/HMM.

and "TOKYO" and "BUILDING" are auxiliary words. A keyword is defined as a word which must be included in an input utterance and for which recognition can be swiftly determined, while an auxiliary word, which may or may not be included in the utterance, is defined as a word which combines with a keyword to construct a compound word.

Next, word HMMs are made for all the keywords and some of the auxiliary words. At the same time, we form connection rules showing which auxiliary words can connect with keywords. The connection rule includes an acceptable connection range, or a number of frame, used in an isolated or overlapped case. In the case of "TOKYO TOSHIBA BUILDING", "TOKYO" is optional before "TOSHIBA", and "BUILDING" is optional after.

After the syntactic analysis of the input word lattice, the final candidate keywords are outputted in order of their likelihoods determined from the normalized scores and the durations of the words for every word sequence (Fig.3).

4 EXPERIMENTAL RESULTS

4.1 SPEECH DATABASE

Three data sets were used for training and evaluating the keyword recognition. The data was sampled at 12 kHz and 16 LPC mel-cepstrum were computed every 8 msec.

The first data set **DS1** was used for designing the codebook in the SMQ. DS1 has 250 phonetically balanced words uttered by 15 male and 15 female speakers. 40,500 segments were manually extracted to design a codebook which includes all 652 Japanese phonetic segments.

The second data set **DS2** was used for training the HMM. DS2 has 583 isolated words uttered by 25 male and 25 female speakers. 32 words and 5 words of DS2 were selected for training the keyword HMMs and auxiliary word HMMs, respectively. The other words were used for training the background HMM.

The third data set **DS3** was used for evaluation. DS3 has 95 compound words including one keyword, uttered by 3 male and 3 female unknown speakers. Five auxiliary words comprise about 1/5 of the non-keywords in DS3.

4.2 KEYWORD RECOGNITION

To examine the effectiveness of the BFP algorithm and the log-likelihood score normalization into word-spotting, we performed a 32-keyword recognition test. The experimental results with the following four algorithms are listed in Table 1:

ALL : All frames are treated as beginning frames;
BFP : The BFP algorithm is used;
ALL-BGN: All frames are treated as beginning frames, and log-likelihood scores are normalized;
BFP-BGN: The BFP algorithm is used, and log-likelihood scores are normalized.

In the word-spotting experiment, the word detected with the highest score is considered to be the keyword. After preliminary tests, the coefficient alpha in equation (3) was set to 0.7.

The recognition accuracy increased when the score normalization by background HMM (BGN) was applied to word-spotting, and a high recognition rate of 87.2% was achieved with a combination of BFP and BGN. Table 1 also shows that the BFP algorithm reduces the computation time by about 90% without decreasing the recognition accuracy.

A recognition experiment on the same data set (DS3) was carried out to evaluate the syntactic analysis of an input word lattice. The vocabulary of the test data includes 32 keywords (building names), 5 auxiliary words, and 17 unknown words. The total number of different words in the data set is 54.

The experimental results are shown in Table 2. Our proposed method achieved the highest recognition score of 91.1% for placing the correct word and the scores within the best 2 and 3 were 96.9% and 98.7%.

5 CONCLUSION

We have examined keyword recognition based on an SMQ/HMM. The recognition accuracy is 87.2% without syntactic analysis, for a 32-keyword recognition task including 5 auxiliary words and 17 unknown words. Our method uses the BFP algorithm which predicts the location of the beginning frame in which the next Viterbi alignment should be generated, and log-likelihood score normalization using a background HMM. Moreover, by incorporating syntactic analysis using connection rules between keywords and auxiliary words, we reached a recognition accuracy of 91.1%.

In future research, we plan to further investigate recognition of spontaneous speech data to handle a large vocabulary and to improve recognition accuracy as steps to constructing a real-time multimodal dialogue system.

REFERENCES

[1] Y.Masai, H.Matsu'ura and T.Nitta, "Speaker Independent Speech Recognition Based on Neural Networks of Each category with Embedded eigenvectors", Proc. ICSLP 90, pp.681-684(1990).

[2] T.Nitta, "Speech Recognition at Toshiba", SPEECH TECHNOLOGY, Vol. 5, Number 4, pp.37-41 (1992)

[3] 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).

[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] R. C. Rose and D. B. Paul, "A Hidden Markov Model Based Keyword Recognition System", Proc. ICASSP90, pp. 129-132 (1990).
[6] H. Tsuboi and Y. Takebayashi, "A Real-Time Task-Oriented Speech Understanding System Using Keyword-Spotting", Proc. ICASSP92, pp. 197-200 (1992).
[7] T. Nitta, T. Murata, H. Tsuboi, T. Kawada and S. Watanabe, "Development of Japanese Voice-activated Word Processor using Isolated Monosyllable Recognition", Proc. ICASSP82, pp. 871-874 (1982).
[8] E. Oja, "Subspace Method of Pattern Recognition", Research Studies Press (1983).
[9] 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).

Table 1: Comparison between word spotting algorithms.

Algorithm	BFP	Background HMM	Recognition Accuracy (%)	Number of Viterbi (ratio)
ALL	None	None	81.6	1.00
BFP	Yes	None	82.5	0.11
ALL-BGN	None	Yes	85.9	2.00
BFP-BGN	Yes	Yes	87.2	0.23

Table 2: Recognition results of compound words with syntactic analysis

Syntactic Analysis	Recognition Accuracy (%)		
	Correct	within the best 2	within the best 3
None	87.2	94.2	97.6
Yes	91.1	96.9	98.7

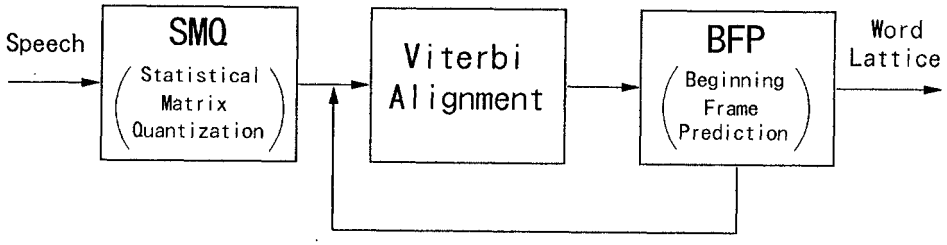

Fig. 2: A block diagram of a word spotting using the BFP algorithm

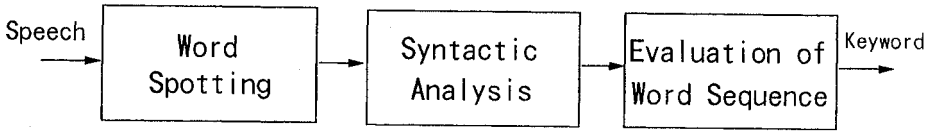

Fig. 3: A block diagram of a keyword recognition with syntactic analysis