WORD-SPOTTING BASED ON INTER-WORD AND INTRA-WORD DIPHONE MODELS

Tsuneo NITTA, Shin'ichi TANAKA, Yasuyuki MASAI and Hiroshi MATSU'URA
Multimedia Engineering Laboratory, TOSHIBA Corporation, Japan

ABSTRACT

In this paper, we propose a precise but simple inter-word diphone model (IDM) for word-spotting based on SMQ/HMM. We have applied ordinary diphone models to a speaker-independent, large-vocabulary word recognition unit. However, because users are apt to add words and/or extraneous speech, accuracy degrades due to the mismatch of models at word-boundaries. The IDM represents transition from the preceding phonemes to a word or from a word to the succeeding phonemes. An experiment showed that the IDMs reduce error rates by about 5% for speech containing unknown words and extraneous speech. The experiment also showed that the proposed method ensured performance good enough for the practical use of a large-vocabulary, isolated-word recognition system.

1 INTRODUCTION

We are developing a speaker-independent. large-vocabulary word recognition unit to use for social information systems with a multimodal user interface (information KIOSKs, ticket-vending machines. ATM. directory assistance systems, etc.)[1]. Because users of these systems sometimes utter spontaneously, the recognition unit needs to be equipped with a word-spotting function. The occurrence frequency of spontaneous speech depends on the application area and user interface design of a system. However, it is important to grasp the system performance tested with mixed data of isolated words (or connected words) and spontaneous speech.

In word-spotting. a filler model is widely used and gives comparatively high recognition performance when the content of speech preceding and/or succeeding key-words is known and these words. phrases. and/or extraneous speech (Uhm, etc.) can be implemented in the filler model beforehand[2]. However. if the spontaneous speech contains unknown words, we can only adopt a phoneme, diphone, or syllable as a filler model[3] and consequently, the word-spotting accuracy degrades.

To cope with the variation at word-boundaries, inter-word context-dependent models were applied to continuous speech recognition systems[4].[5]. In this paper, we propose a precise but simple inter-word diphone model (IDM) for word-spotting based on SMQ/HMM[6]. Because we must implement the speech recognition unit to various types of social information service systems. re-designing of the unit is resource-consuming and undesirable. The proposed method requires no additional speech data and training either for inter-word models and for intra-word models.

This paper is organized as follows. Section 2 describes the word-spotting algorithm based on SMQ/HMM. Section 3 explains the proposed model and its implementation into keyword models. Section 4 shows the results of experiments on a large vocabulary task using both an isolated word database and spontaneous speech database.

2 SMQ/HMM-BASED WORD-SPOTTING

2.1 SMQ/HMM

The recognition system based on SMQ/HMM[6] is roughly divided into two stages: a phonetic segment decoding stage that performs statistical matrix quantization (SMQ), and a word matching stage that uses the Viterbi algorithm. We incorporated detailed phonetic variations less than 100 msec. in duration into an orthogonalized phonetic segment codebook of SMQ. and speech variations more than 100 msec. in duration into word HMMs. The block diagram of the basic system is shown in Figure 1.

Many variations are observed in continuous speech. Some of them can be described only by VCV combination and others by acoustic segment. Therefore. we need to use multiple phonological structure elements to describe speech. A phonetic segment[6] extracted from a Japanese speech database consists of about 652 acoustic and phonetic structures in total varying in duration from 32 to 96 msec. (e.g.. acoustic segments, phonemes $(C\ and\ V).C_V.CC,_{VC}\ and\ {}_VC_V)$.

The SMQ method effectively incorporates the pattern variations of each phonetic segment into the orthogonalized phonetic segment 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 a phonetic segment c and a normalized input pattern $X_i = (x_{11},\dots x_{nt},\dots x_{NT})$ 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. Equation 1 is the same as the expression used in the Multiple Similarity Method[7] and the Sub-space Method[8].

The HMM handled in this paper is a left-to-right model of a discrete density HMM. Transition probabilities and output probabilities are calculated using K-best codes in the SMQ method[9] and estimated by the forward-backward algorithm[10]. The optimal state sequence in HMM networks is searched using the Viterbi algorithm. The SMQ/HMM has achieved a high performance in a speaker-independent and large-sized vocabulary word recognition tasks[9].

2.2 Subword Model and Filler Model

The word-spotting algorithm that we developed uses word-HMMs and filler-HMMs[3]. Any word-HMM can be formed with some of the 235 discrete sub-word HMMs in diphones [11]. A diphone model consists of 3 loops and 4 states, and each state consists of 652 phonetic segmnet output probabilities and transition probabilities. On the other hand, filler-HMMs can be formed with 100 Japanese mono-syllable models. Mono-syllables are classified into 2 types by phonetic structure, V and CV. The V model consists of 3 loops and 4 states, and the CV model 6 loops and 7 states. Figure 2 shows a model in which filler-HMMs can express any string of syllables and come before and after each of word-HMMs. A penalty is given to the each output probability of the filler-HMMs to avoid matching the filler-HMMs with a keyword speech. This model can handle any speech that include extraneous speech. In addition, for speech having only a word and no extraneous speech, the model can work the same in matching performance as a word-HMM with no filler. Therefore, it is close to a conventional word model in performance.

3 INTER-WORD DIPHONE MODEL

In isolated word recognition, we used a diphone model that represents transition from the preceding silence to a phoneme or transition from a phoneme to the succeeding silence as a word-boundary diphone model. However, word-spotting accuracy degrades due to the mismatch of models at word-boundaries. We propose an "Inter-word diphone model (IDM)" that represents transition from the preceding phonemes to a word or from a word to the succeeding phonemes. Since the Japanese language is open-syllabic in structure, the IDM has only 7 preceding phonemes (/a/, /i/, /u/, /e/, /o/, /Q/ and /N/) where /Q/ and /N/ are silence and independent nasal sound, respectively. That is, any Japanese word begins with one of the 7 preceding diphone models. The Output probability B_i^w of the ith state of a diphone model (intra-word diphone model) w is represented as follows:

$$B_i^w = (b_{i1}^w, b_{i2}^w, \dots, b_{iL}^w) \tag{2}$$

where L denotes the number of phonetic segments. A preceding inter-word diphone model (PDM) is designed to represent the most suitable model among the 7 preceding diphone models and is defined as follows:

$$B_1^{PDM} = (\max_{w \in W_P} b_{11}^w, \max_{w \in W_P} b_{12}^w, \dots, \max_{w \in W_P} b_{1L}^w) \tag{3}$$

and,

$$B_i^{PDM} = (b_{i1}^P, b_{i2}^P, \dots, b_{iL}^P) \qquad (2 \le i \le 3) \tag{4}$$

where W_P is a set of preceding diphone models and P is an intra-word diphone model with preceding silence.

The transition probabilities of the PDM are equal to those of the intra-word diphone model P except for the first self-transition probability. The first self-transition probability is made smaller than that of P.

In the same way, a word ends with one of the 20 succeeding diphone models. A succeeding inter-word diphone model (SDM) is defined as follows:

$$B_i^{SDM} = (b_{i1}^S, b_{i2}^S, \dots, b_{iL}^S) \qquad (1 \le i \le 2) \tag{5}$$

and,

$$B_3^{SDM} = (\max_{w \in W_S} b_{31}^w, \max_{w \in W_S} b_{32}^w, \dots, \max_{w \in W_S} b_{3L}^w) \tag{6}$$

where W_S is a set of succeeding diphone models and S is an intra-word diphone model with succeeding silence. The last self-transition probability is made smaller than that of S.

For example, the word "Tokyo" has a PDM for preceding diphones (/Qt/, /at/, /it/, /ut/, /et/, /ot/ and /Nt/) and SDM for suceeding diphones (/oQ/, /oa/, /oi/, /ou/, /oe/, /oo/, /ok/, /os/, /ot/, /on/, /oh/, /om/, /oj/, /or/, /ow/, /og/, /oz/, /od/, /ob/ and /op/). The PDM and SDM are used in place of the /Qt/ and /oQ/ diphone models in the word "Tokyo", respectively.

4 EXPERIMENTAL RESULTS

Data sets D-1 and D-2 were used for training and other data sets D-3 and D-4 were used for evaluation.

1. D-1 was used to train the codebook of phonetic segments by SMQ. D-1 includes 250 phonetically balanced words uttered by 15 male and 15 female speakers. Out of these words, 40,500 segments were manually extracted to train the codebook that includes 652 full Japanese phonetic segments.

2. D-2 was used to train sub-word-HMMs and filler-HMMs, and consists of 492 isolated words that include all the Japanese VCV contexts. Each of 20 male speakers uttered these words one time each.

3. D-3 was used for word-spotting evaluation. D-3 consists of 227 isolated words used in the directory assistance system. Each of 4 male speakers, other than those for D-1 and D-2, uttered the words one time each.

4. D-4 was used also for word-spotting evaluation. D-4 consists of 73 keywords selected out of the 227 words in D-3. However, the keywords are preceded by either of the interjections, "eeto", "eh" and "anoh", (expressions similar to "well..." in English), and followed by either of the short phrases, "Onegaishimasu" ("please"), "wadokodesuka" ("where is"), "wo oshiete" ("tell me") and so on. Each of 4 male speakers, who are the same as those for D-2, uttered once each of the keywords preceded by one of the interjections and followed by one of the short phrases. D-4 data is hereinafter reffered to as "non-isolated words".

Figure 3 shows the results of recognition experiment on data sets D-3 and D-4. using and without using the IDMs that we proposed in this paper. The results are given in recognition error rates for 227 words at various penalties on the output probabilities of filler-HMMs (the penalty is hereinafter referred to as "filler-penalty"). In the experiment using the IDMs, penalty on the PDM's self-transition probability in the first state and that on the SDM's self-transition probability in the third state (these two penalties are hereinafter referred to as "PDM-penalty" and "SDM-penalty". respectively) are set at infinity. That is, the states have no self-transition.

Increasing the filler-penalty increases the likelihood of word-HMMs relatively to that of filler-HMMs. thus higher filler-penalty gives a lower error rate as shown in Figure 3 when recognition is made on D-3, the data set of isolated words.

On the other hand, a proper filler-penalty minimizes the error rate when recognition is made on D-4, the data set of non-isolated words. This is because word-HMMs more easily spot key words since filler-HMMs match unknown words and extraneous speech properly. Figure 3 reveals that using the IDMs improves recognition performance by about 5%.

Figure 4 shows experimental results for 227 words at various PDM-penalty values and SDM-penalty values. The results shows that the recognition performance of the proposed method greatly improves when increasing the PDM-penalty and the SDM-penalty. The filler-penalty values used for the data in the figure are 500. 1000 and 2500. The figure reveals that a filler-penalty of about 1000, and PDM-penalty and SDM-penalty of about 2000 are most suitable.

Figure 5 shows experimental results that change filler-penalty, PDM-penalty. and SDM-penalty to achieve the best recognition performance. The experiment is made on data sets D-3 and D-4 for 227 words to be recognized. and for 500 and 1000 words to be recognized both including the 227 words. The horizontal axis represents the percentage of non-isolated words in the data given to the system. The users of a speech recognition system are apt to add words and/or extraneous speech when they speak. Figure 5 gives the error rate of a word recognition system based on the proposed method for any percentage k of unknown words and extraneous speech in the input speech. For example, a non-isolated word percentage k of 10% gives an error rate of about 5% when a 1000-word set is used. and even that of 20% gives an error rate of about 6% when the number of vocabulary is limited to 227. The percentage(k) depends on the task and user interface of a system. However. since our experience estimates that the percentage falls in a range of 10 to 20% when the speech recognition unit is applied to social information services. the results shown above tell that the proposed method gives enough performance good enough for practical use.

5 CONCLUSION

We examined in various ways a word-spotting algorithms using filler-HMMs. The examination aimed at alleviating the degradation of recognition performance caused by unknown words and extraneous speech that users add in their utterances. The IDMs and penalties that we proposed in this paper reduced error rates by about 5% for speech containing unknown words and extraneous speech. The experiments revealed that a non-isolated word percentage of 20% or less allowed recognition performance good enough for practical use.

References

[1] H.Matsu'ura.Y.Masai. J.Iwasaki. S.Tanaka. H.Kamio and T.Nitta, "Multimodal, Keyword-based Spoken Dialogue System MultiksDial" Proc. ICASSP94, Vol.2. pp.33-36, 1994.

[2] J.G.Wilpon, L.R.Rabiner, C.-H.Lee and E.R.Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", IEEE Trans. on Acoustics and Signal Processing, Vol.38, No.11, pp.1870-1878, 1990.

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

[4] W.Chou, T.Matsuoka, B.-H.Juang and C.-H.Lee, "An Alogorithm of High Resolution and Efficient Multiple String Hypothesization for Continuous Speech Recognition using Inter-word Models", Proc. ICASSP94, Vol.2. pp.153-156, 1994.

[5] A.Nakamura. "A Minimum Error Training of Garbage Model for Keyword Spotter with Artificially Generated Training Data". Proc. EUROSPEECH 95. Vol.3. pp.1641-1644. 1995.

[6] 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.

[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, pp871-874. 1982.

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

[9] 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.

[10] 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.

[11] Y.Masai,J.Iwasaki,S.Tanaka,T.Nitta,M.Yao,T.Onogi and A.Nakayama. "A Keyword-Spotting Unit for Speaker-Independent Spontaneous Speech Recognition". Proc. ICSLP94, pp.1383-1386, 1994.

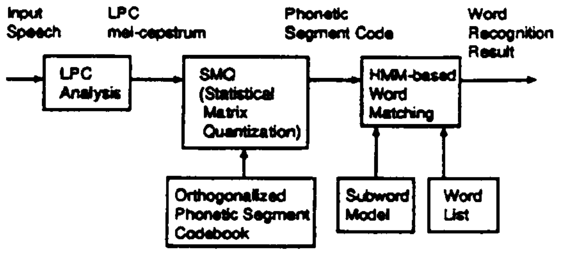

Figure 1: Block diagram of the basic speech recognition system

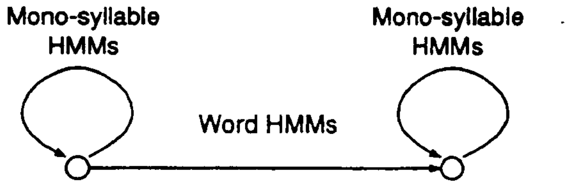

Figure 2: Word-spotting with the filler of mono-syllable HMMs

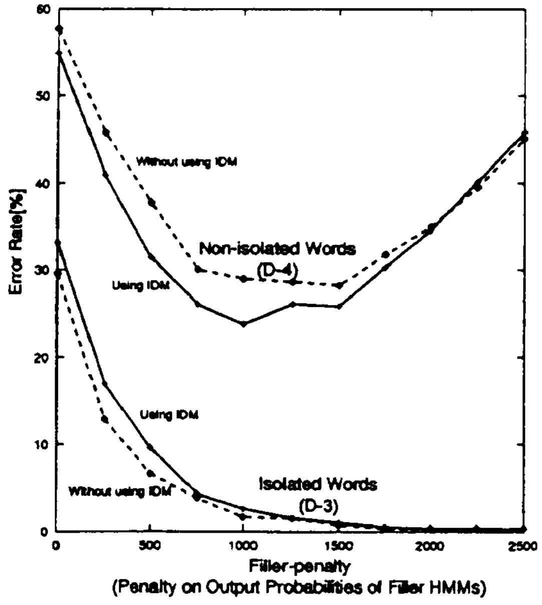

Figure 3: The recognition error rate using and without using IDMs for 227 words(D-3. D-4) at various filler-penalty values

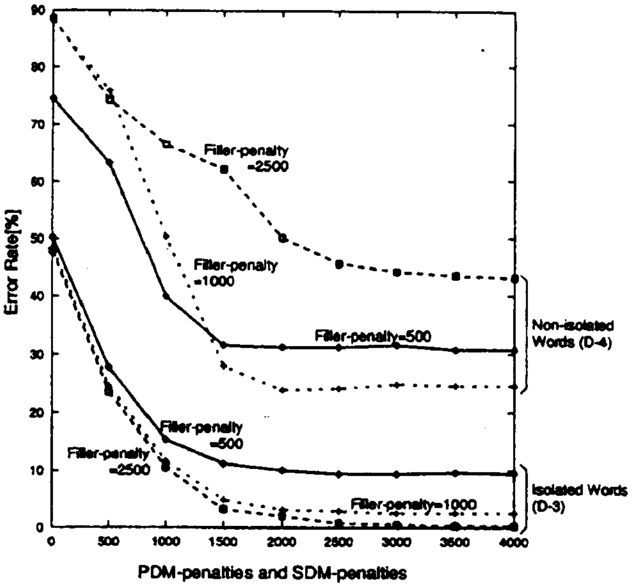

Figure 4: The recognition error rate using IDMs for 227 words(D-3, D-4) at three filler-penalty values(500, 1000, 2500) and at various PDM-penalty values and SDM-penalty values

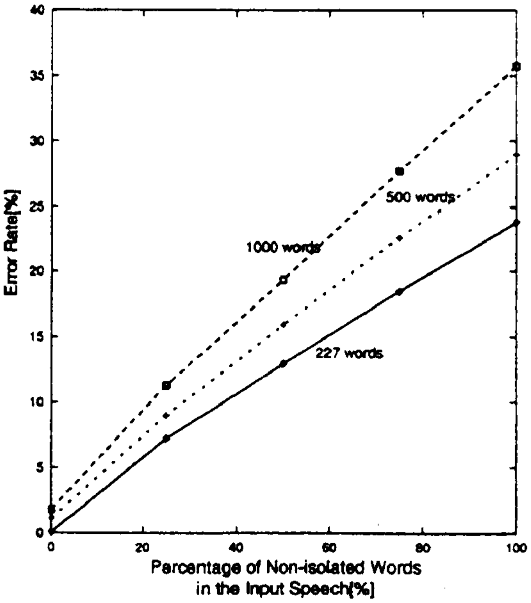

Figure 5: The error rate for 227, 500 and 1000 words recognition system based on IDMs for any percentage of non-isolated words in the input speech