A MEMORY MANAGEMENT METHOD FOR A LARGE WORD NETWORK

T.Iwasaki and Y.Abe
Human Media Technology Dept.
Information Technology R&D Center
MITSUBISHI Electric Corp.
5-1-1, Ofuna, Kamakura, Kanagawa, 247, Japan

ABSTRACT

To improve the performance of continuous speech recognition, it is effective to incorporate grammatical knowledge of task into a word network of a FSN (finite state network) form. But, recently, some of them requires huge memory, so we introduce an efficient memory management method for a large word network; distributed FSN model and hiearchical memory model. The system keeps the word network divided to small sub-networks, and activates each sub-network when necessary. Using this method, we can recognize continuously spoken sentences of Japanese addresses, which are made of 390K geographic names, with only 5.6 Mbytes local memory in average.

1.INTRODUCTION

This paper describes a memory management method of a large word network represented by the finite state network (FSN). To improve the performance of continuous speech recognition, it is effective to incorporate grammatical knowledge of task into a word network in a FSN form. Needs to handle a huge word network have already been raised in recognition of address, for example, and the knowledge to be handled would become huge as number of fields is rapidly increasing where the speech recognition is put in practical application. For the purpose, effective memory management method is required for practical speech recognition devices. For the speech recognition system with multi-processor, in particular, it is important to manage the working memory for recognition process as separated from the word network since local memory of individual processors is restricted from the aspect of cost. A number of systems has been proposed, on the other hand, which perform recognition function by converting to FSN dynamically activated from the grammatical rule described by the context free grammar (CFG) [1][2]. CFG, however, is inadequate to handle such a large scale word network as the address. Accordingly, the authors will introduce the distributed FSN model and a hierarchical memory model which dynamically activates the sub-network when necessary. This model also has solved the problem of dynamic generation of context-dependent model.

2.MEMORY MANAGEMENT METHOD

2.1 Address database

The task is recognition of addresses throughout Japan. The database of addresses in Japan [3] is used as the source of knowledge for the task. The database contains 390K geographic names (number of different words is 210K) from names of prefectures to those of sub-towns and information of chome, which is a section of town or sub-town, together with the link information among these names.

2.2 Distributed FSN model

The model recognizes the speech of addresses in Japan from names of prefectures to numbers of blocks and buildings. Figure 2 shows a word network structure to represent these addresses. The knowledge of address database is embedded from prefectural names to chome.

As the above word network has a size of several tens Mbytes, it is divided to sub-networks of small size. The

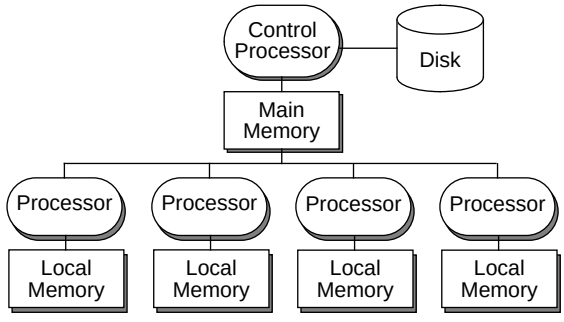

Figure 1:Speech recoginizer with multi-processor

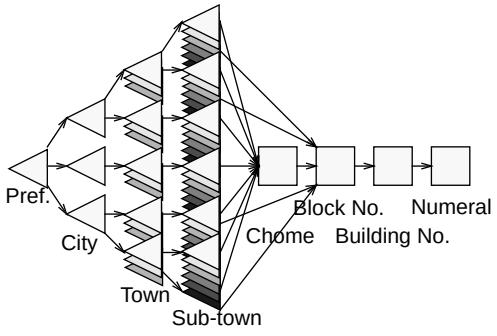

Figure 2:Word network to represent addresses of Japan

sub-network is called as the context container (referred to as CC hereinafter). For address recognition, compact units such as name of city in a prefecture are used for CC.

We will explain the structure of CC. Figure 3 shows the example of addresses used for explanation. They are not actual addresses, and they are shorter than actual addresses. A word enclosed in the parentheses { and } indicates it may be omitted. Figure 4, shows the CC, representing the addresses shown in Figure 3. CC is composed of node shown by box and arc shown by arrow. The node mainly represents information of word. Three types of node exist: (1)Entry node, (2)word node and (3)connection node. In C3 of CC in Figure 4, for example, entry node is E1, word nodes are W1, W2 and W3, and connection nodes are C5:E1 and C5:E2.

- Nishi-machi{Aza}Shimokawa 1-chome
- Nishi-machi{Aza}Uenoyama 1-,2-chome
- Higashi-machi{Aza}Kamisato 1-,2-chome
- Higashi-machi{Aza}Shimomura 1-,2-,3-chome

Figure 3:Example of address for explanation

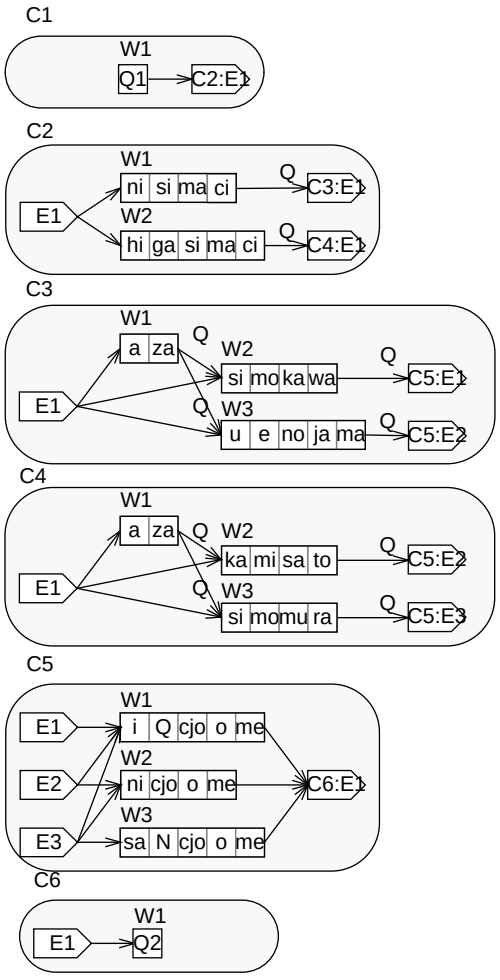

Figure 4:Structure of context container

The entry node represents position where other CC can be connected, and may have more than one entry nodes like C5. The word node contains syllable description of word and has attributes of word and numeral. By difference of these attributes, the same reading of dif ferent HMM model can be used. The connection node means connection to other CC. For example, C5:E1 means connection to the entry node of E1 of C5. On the other hand, each arc has attribute to indicate whether or not insertion of pause and filled pause is allowable. The attribute of Q is given in Figure 4, which indicates insertion of pause. As shown by this example, CC is a model of distributed FSN expression.

2.3 Hierarchical memory structure

A recognition engine uses the hierarchial memory structure for recognition processing while loading this CC when necessary. Figure 5 shows this structure. Set of CC is stored in area called as container warehouse. This area is assigned to the main memory or disk of the control processor. When the recognition processing is started, the recognition engine loads CC from the container warehouse when necessary and deploys on the local memory of recognition engine. The local memory is divided to three areas: (1)Container shelf, (2)Viterbi work memory and (3)back track memory. The container shelf is a memory to store loaded CC in the data structure which is adequate to refer. The Viterbi work memory is a work area used for V iterbi calculation. In the back track memory, data is stored for N-best search to obtain final results of search. Technique of CC deployment is described below.

First, when the recognition engine requires new CC that is not found on the container shelf, it loads CC from the container warehouse, converts data structure, and stores on the container shelf. The required memory is enlarged to about 5 times of memory volume at this stage. CC once stored in the container CC is placed in the memory and used again when the recognition engine requires to connection it. Then, the recognition engine refers CC on the container shelf and allocates data area for a syllable HMM required for V iterbi calculation. This area is called as the V iterbi work mem-

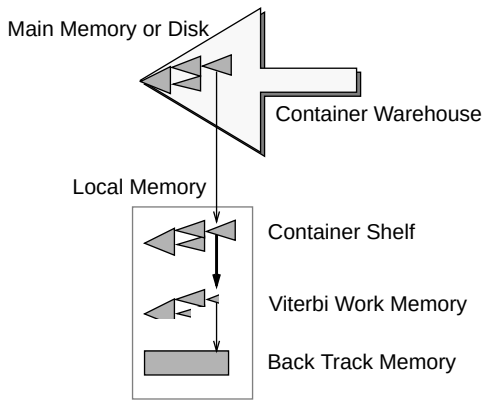

Figure 5:Hiearchcal memory model

ory. At this time, the attribute of arc is evaluated and pause and filled pause are automatically inserted between nodes. Figure 6 shows content of the V iterbi work memory, when “Nishi-machi {Aza} Shimokawa” is spoken. We assume that C1, C2 and C3, which are CC that contain these spoken geographical names, are downloaded on the container shelf from the container warehouse.

Arcs between “Nishi-machi” as C2 and “Aza”, “Shimokawa” and “Kaminoyama” as C3 have attributes that allows insertion of pause. In C3, pause can be inserted between {Aza} and “Shimokawa” or “Kaminoyama”. We use the syllable HMM model that depends upon preceding phoneme. But some of them depend upon succeeding phoneme. They are activated on V iterbi work memory, when necessary under consideration of preceding or succeeding phoneme. For example, “Q/ni” is the model of syllable “ni”, having preceding phonetic environment “Q”. The network grows with the phonetic environment taken into consideration. At this time, only the branch of “Nishi-machi” grows in C2 and the branch of “Higashi-machi” which is different from the speech does not grow at all since a branch of syllable model is allowed to grow only when its distance value of final state is smaller than a threshold. If the attribute of arc allows to insert a pause, word nodes of pause, “i/Q3” and “a/Q3” are generated and inserted between word nodes of geographic names. If there are more than one preceding word nodes, and their phonetic environment differs to each other, the first syllable is separated at each of phonetic environment and a syllable network of adequate phonetic environment is generated. Data for backtracking is output to the back track memory between the syllable models as a principle. Data is output only if the distance value of final state of the preceding syllable model is smaller than a threshold. Data is not output, however, to the back track memory without having plural preceding syllable model, since unambiguous backwarding is possible at the time of back tracking. The black triangle in Figure 6 show the points that output data to back track memory. Even when the number of preceding syllable model is one at the time of its generation, connection may be added later from more than one preceding syllable models since the syllable network gradually grows. Therefore, data is output to the back track memory starting when it has more than one preceding syllable models and continuing until the calculation is terminated by beam search.

3.RECOGNITION EXPERIMENT

3.1 Speech data

Table 1 shows speech data for evaluation used in the experiment. It is addresses spoken through telephone line according to the guidance of automatic recording system. Some of data contain background noise of residence or office (sound of TV or ringing bell, conversation behind the speaker) or it may often be saturated when recording. Speeches were excluded when they were mistaken, contained name of building or filled pause as they cannot be accepted by the word network of address. Speech section is manually picked up from the speech data.

Table 1 :Speech Data

Input devices	Telephone set
Location of utterance	25 Prefectures in Japan
Location of recording	Tokyo
No. Speaker	344

3.2 Experimental conditions

Table 2 shows analytical conditions applied to the experiment of recognition. Table 3 shows HMM used for it. The acoustic phonetic segment (APS) is basic distribution unit of HMM which is smaller than the phoneme. It is basically the model of a state that depends upon one or two preceding phoneme. Each syllable of Japanese language is composed of from 2 to 7 APS’ s. For recognition of geographic names, 327 forms of APS’s are used, whereas 401 forms of APS’ s are used for recognition of numeric’s. The former is trained from the word corpus of different phonetic balance from geographic names and the latter is trained from speech data that contains a large variety of speeches of numerals. A total of 22,915 CC was prepared from the address database described above. The volume was 42 Mbytes on the container warehouse. The recognition experiment was conducted with the rule to allow insertion of pause between geographic names as the attribute of arc.

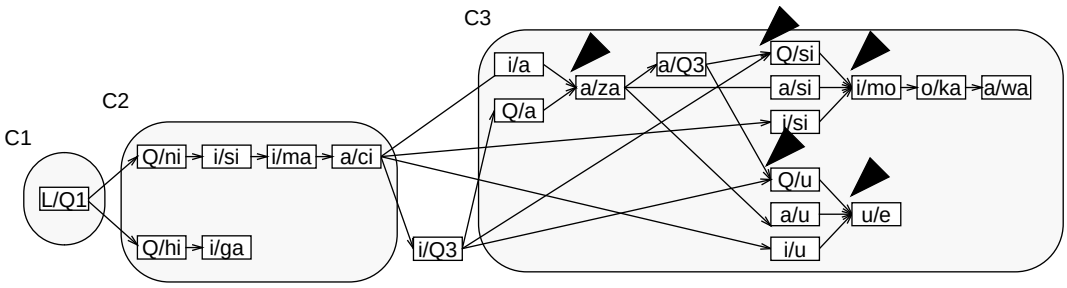

Figure 6:Syllable network on Viterbi work memory

Table 2 :Analysis conditions

Sampling freq.	8kHz
Filter	300Hz ～ 3.4kHz
Pre-emphasis	$1\text{-}0.95z^{-1}$
Frame period.	10msec
Frame length	25msec
Parameter	mel-cepstral features ＋ delta-cepstral features ＋ delta-power(21 dimensions)

Table 3 :HMM

HMM type	Continuous HMM
Recognition unit	Acoustic Phonetic Segment (APS) [4] (728 kinds、1 state per APS)
Distributions	Diagonal Gaussian distribution
No. mixture	7
Duration	Univariance Gaussian distribution
Viterbi	Simplified duration control[4]-Frame synchronous beam search

3.3 Experimental result and review

Figure 7 shows results of the experiment. Recognition experiment was conducted in varying thresholds. The thresholds are concerned in beam search, transition to next syllable model, and output backtracking data. Results are plotted for the volume of local memory on the horizontal axis and the sentence accuracy on the vertical axis. The memory for speech analysis and calculation of output probabilities is not included in memory volume in Figure 7. The volume of memory is converted to volume with 32 bits CPU, which means that the pointer size is 4 bytes. The sentence accuracy decreases as the volume of memory is reduced. The volume of local memory was kept at about 5.5 Mbytes in average when the sentence accuracy was 53% (geographic names: 91.9 %; Chome and block Nos.:52.3%). The volume of memory is broken down as shown in Table 4.

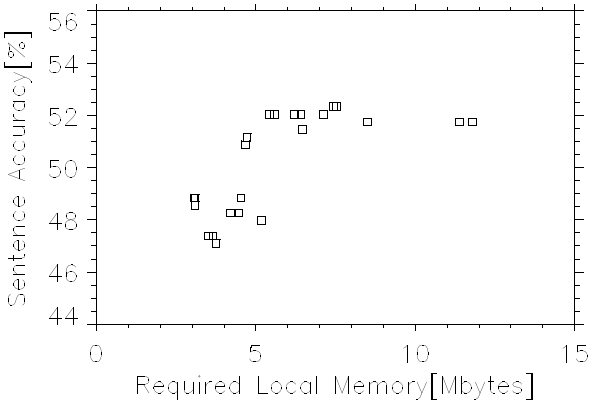

Figure 7:Memory volume vs. sentence accuracy

Table 4 :**Volume of memory of local memory**

Container Shelf	0.7Mbytes
Viterbi Work Memory	2.1Mbytes
Back Track Memory	2.7Mbytes

Only 0.2% of total number of CC was downloaded to the container shelf. Moreover, of the syllables downloaded on the container shelf, 29.5% was activated on the Viterbi work memory. Whereas most of the container shelf and Viterbi work memory is consumed by recognition of geographic names. 70% of the back track memory was used by the chome and Nos. This is because most of branching to backward exist on the chome and Nos. segments. The volume of back track memory would be significantly reduced by calculation with 1-best instead of N-best. The volume of memory often explodes and in certain cases a local memory of maximum 25 Mbytes was required for speeches to be accepted as correct. Since a significant volume of local memory was not used by the beam search, the memory volume may be further saved releasing unused memory area.

As the syllable HMM model is variable length of states, the overhead time for memory management was elongated to an unacceptable level, when allocation and release are often repeated. According to this judgement, local memories are not released after they are once secured in the present experiment. This has led to continuous increase of local memory volume until the termination of speech. Since it is possible to employ a technique to release local memory altogether when the speech is paused, significant reduction of memory volume is likely to be achieved without the increase in the calculation amount.

4.CONCLUSION

The author examines a memory management method to recognize while deploying the distributed FSN in a hierarchial memory model. Its ef fectiveness was affirmed by an experiment of recognition of addresses throughout Japan. Studies would be continued on technique to control the explosion of memory volume and method of garbage collection.

REFERENCES

[1] M.K.Brown,S.C.Glinski,"Context-Free Large Vocabulary Connected Speech Recognition With Evolutional Grammars", ICASSP94,pp.II.145-148(1994.4)

[2] K.Kita,Y.Yano,T.Moriyama,"One-Pass Continuous Speech Recognition Directed by Generalized LR Parsing", ISCLP,1.4,pp.13-16(1994)

[3] KOKUDOCHIRIKYOUKAI ZENKOKU MACHI AZA File

[4] T.Iwasaki,K.Nakajima,"A Real Time Speaker-Independent Continuous Speech Recognition System",-ICPR92,pp.663-666(1992.9)