A JAPANESE TEXT DICTATION SYSTEM BASED ON PHONEME RECOGNITION AND A DEPENDENCY GRAMMAR

Shozo Makino, Akinori Ito, Mitsuru Endo and Ken'iti Kido

Research Center for Applied Information Sciences,
Tohoku University Sendai, 980 Japan.

ABSTRACT

A Japanese text dictation system has been developed based on phoneme recognition and a dependency grammar. The phoneme recognition is carried out using a proposed modified LVQ2 method. Phoneme recognition score was 86.1% for 218 sentences uttered by two male speakers. A linguistic processor is composed of a Bunsetsu-unit(Japanese Phrase) spotting processor and a syntactic processor with semantic constraints. The structure of the Bunsetsu-unit is effectively described by a finite state automaton. The perplexity of the finite state automaton is 230. In the Bunsetsu-unit spotting processor, using a syntax driven continuous DP matching algorithm, the Bunsetsu-units are spotted from a recognized phoneme sequence and then a Bunsetsu lattice is generated. In the syntactic processor, the Bunsetsu lattice is parsed based on the dependency grammar. The dependency grammar is expressed as the correspondence between a FEATURE marker in a modifier-Bunsetsu and a SLOT-FILLER marker in a head-Bunsetsu. Recognition scores of the Bunsetsu-unit and conceptual word were 73.2% and 85.7% for the 218 sentences uttered by the two male speakers.

I. INTRODUCTION

A number of continuous speech recognition systems have been reported. However there still remain several problems in developing a continuous speech recognition system for ordinary Japanese text utterances, because the traditional continuous speech recognition systems only dealt with particular linguistic information in a specified task domain. In order to build up a Japanese text dictation system, we should develop the following methods:

(1)Phoneme recognition method with high accuracy
(2)Bunsetsu-unit spotting method with high accuracy and with a small amount of computation
(3)Efficients parsing method taking into account syntactic and semantic constraints

This paper proposes a modified LVQ2 method for the phoneme recognition, a syntax driven continuous DP for the Bunsetsu-unit spotting and a CYK based parsing method for the syntactic processing. Finally this paper describes a proto-type system which can convert continuous speech uttered Bunsetsu by Bunsetsu to a Japanese Kanji-Kana string.

II. OUTLINE OF THE JAPANESE TEXT DICTATION SYSTEM[1-7]

Figure 1 shows a schematic diagram of the Japanese text dictation system. The system is composed of an acoustic processor[1], a Bunsetsu-unit spotting processor[2-4] and a syntactic processor with semantic constraints[5-7]. In this research the speech to be recognized is spoken sentences which are syntactically and correct. We use sentences from a scientific paper, where the sentences have 843 conceptual words and 433 functional words. Average numbers of Bunsetsu-units and phonemes in a sentence are 4.1 and 47.

Input speech is analyzed by a 29 channel band-pass filter bank. In the acoustic processor a phoneme sequence is recognized from the input speech using a proposed modified LVQ2 method[1].

The structure of the Japanese sentence is effectively described by a two-level grammar which consists of an intra-Bunsetsu grammar and an inter-Bunsetsu grammar. Thus analysis of the Japanese sentence is divided into two stages. The first stage is the extraction of the Bunsetsu-unit candidates from the recognized phoneme sequence. The second one is the analysis of the dependency structure among the Bunsetsu-unit candidates. The Bunsetsu-unit is composed of a conceptual word and several functional words. The Bunsetsu-unit can be represented by a finite state automaton.

III. PHONEME RECOGNITION USING A MODIFIED LVQ2 METHOD[1]

Learning Vector Quantization(LVQ,LVQ2) methods were proposed by Kohonen et al.[9]. McDermott et al.[10] developed a shift-tolerant phoneme recognition system based on the LVQ2 method. We propose a

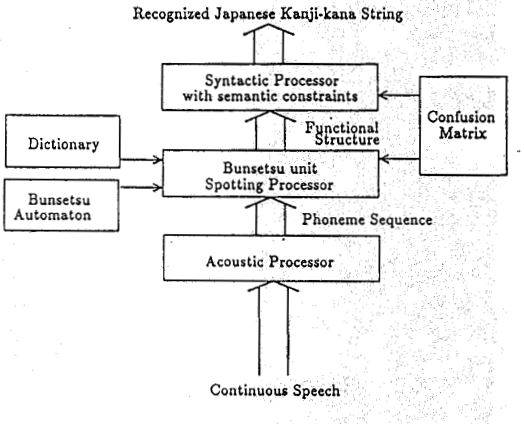

Figure 1 A schematic diagram of the Japanese text dictation system.

CH2977-7/91/0000-0273 $1.00 © 1991 IEEE

modified training method for the LVQ2 method. In the LVQ2 algorithm proposed by Kohonen, two reference vectors are modified at the same time if the first nearest class to an input vector is incorrect and the second nearest class to the input vector is correct. In the modified LVQ2 algorithm, n reference vectors are modified at the same time if the correct class is within the pre-defined N-th rank.

Figure 2 shows the training method of the modified LVQ2 algorithm. In step 1, reference vectors are chosen using the K-Means clustering method from each class. In step 2, the nearest reference vector of each class to an input vector is selected. In step 3, the rank of the correct class is computed. When the rank of the correct class is n, we assume that the reference vector of the correct class is m_n. In step 4, n is checked to see whether or not n falls in the range of $2 \leq n \leq N$. In step 5, whether or not the input vector falls within a small window is checked, where the window is defined around the midpoint of m_1 and m_n. In step 6, the i-th reference vector is modified according to the following equations.

$$[m_i]^{t+1} = [m_i - \alpha(n)(x - m_i)]^t$$
$$(i = 1, 2, ---, n-1)$$
$$[m_n]^{t+1} = [m_n + \alpha(n)(x - m_n)]^t$$

The phoneme recognition system is similar to the shift-tolerant model proposed by McDermott et al.[10]:

(1) 8 cepstrum coefficients and 8 Δ cepstrum coefficients are computed every frame from the 29 channel BPF spectrum. Each reference vector is represented by 112 coefficients(7 frames x 16 coefficients). Each class was assigned 15 reference vectors chosen by the K-Means clustering method.

(2) A 7-frame window is moved over the input speech and yields a 112(16x7) dimensional input vector every frame.

(3) In the training stage the modified LVQ2 method is applied to the input vector as described above.

(4) In the recognition stage we compute distances between the input vector and the nearest reference vector within each class.

(5) From this distance measure, each class was assigned an activation value a_w as follows:

$$a_w(c,t) = 1 - d(c,t) / \sum_i d(i,t)$$

where d, c and t are distance, class and time.

(6) Final activation a_f is computed by the following equation.

$$a_f(c,t) = \sum_{j=-4}^{4} a_w(c,t+j) g_w(j)$$

The final activation is defined by summing 9 activation values. g_w is weight of the gaussian window.

(7) A class with maximum activation value is regarded as phoneme candidate of each frame. The activation value is regarded as an a posteriori probability $P(C_k|t_k)$ of the phoneme C_k at the t_k-th frame.

(8) An optimum phoneme sequence is computed from the phoneme candidate sequence using a dynamic programming and duration constraints[8].

IV. BUNSETSU-UNIT SPOTTING PROCESSOR[2-4]

There are two traditional methods for extraction of the Bunsetsu-units. The first method(method 1) spots the Bunsetsu-units using all possible Bunsetsu-unit reference patterns. However, the method needs a large amount of storage and computation because the number of Bunsetsu-units is huge in the Japanese text dictation system. On the other hand, the second one(method 2) spots the conceptual words and the functional words independently. However, the Japanese language has many functional words with short lengths such as copulae, inflection of conjugated words and auxiliary verbs. The present spotting method shows poor performance in spotting short length words and therefore insertion and deletion errors are common although the amount of computation for this method is very small. The method(method 3) proposed in this paper is an intermediate one. This method spots the Bunsetsu-units based on the finite state automaton representing the Japanese Bunsetsu-unit structure. We call this method a syntax driven continuous DP.

The storage capacity necessary for storing the finite state automaton is greatly reduced comparing to the method 1. Also the computation amount is much reduced using the continuous DP and a beam search. Figure 3 shows the outline of the structure of the Bunsetsu-unit. The arcs represent adverbs, adnominals, nouns, verbs and adjectives etc. Double circles in the figure are terminal states. The perplexity of the finite state automaton is 230.

Figure 4 shows an example of processing with the syntax driven continuous DP. The processing starts with a conceptual word. If the likelihood at the final phoneme of the stem of the conceptual word exceeds a threshold, the automaton generates the next words. The calculation of likelihood for the next word is carried out using the final likelihood obtained at the previous stage as the

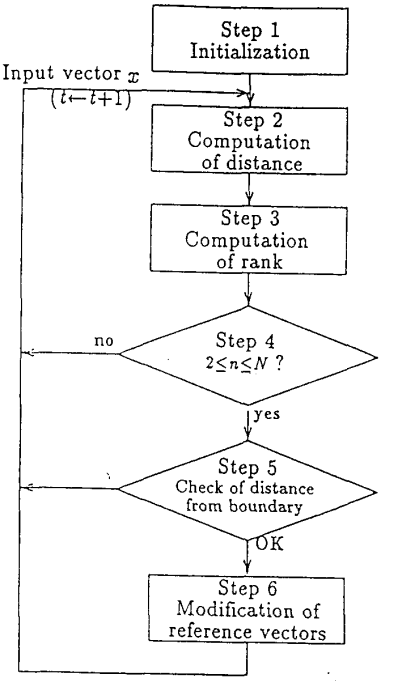

Figure 2 Algorithm of modified LVQ2 method

initial value. In the same manner the calculation of likelihood continues until each path reaches its terminal state. If the likelihood at each of a terminal state exceeds a threshold, the Bunsetsu-units of every valid path to that terminal state are recognized as candidates and thus a Bunsetsu lattice is made from an input phoneme sequence. Using the syntax driven continuous DP the Bunsetsu-units are spotted from the input phoneme sequence and simultaneously the morpheme analysis is carried out.

Figure 5 shows the relation between number of candidates per 100 input phonemes and the detection score in spotting conjugated words. Method 3 shows similar performance to method 1. Method 2 detects 20 times more candidates compared to method 3 when the detection score is 90%.

V. SYNTACTIC PROCESSOR WITH SEMANTIC CONSTRAINTS[5-7]

Syntactic processing is applied to the candidates of the Bunsetsu-unit in the Bunsetsu lattice. When the Bunsetsu-unit is detected, a functional structure for a syntactic processor is created. This structure corresponds to a partial tree of phrase structure tree. Most features in a functional structure are given from lexical items, where the features contain syntactic and semantic information. The inter-Bunsetsu grammar is implicitly expressed as the correspondence of markers in functional structure of Bunsetsu-unit candidates. Two partial trees are merged when the modifier's FEATURE marker set and the head's SLOT-FILLER marker set contains same syntactic marker. We use 95 syntactic and semantic features. All syntactic dependency, including modification, complement, object and subject are treated in the same framework. Figure 6 shows an example of the merging of two Bunsetsu-units. In this example the Bunsetsu-unit A has a FEATURE "a" and the Bunsetsu-unit B has a SLOT-FILLER of the feature "a". Thus the Bunsetsu-units A and B are merged.

The algorithm for parsing is based on the C.Y.K. algorithm using the beam search. This method can give multiple results of the sentence for the input phoneme sequence. The computation amount is $O(N^3D^2)$, where N is the length of an input phoneme sequence and D is the number of the candidates stored in each interval.

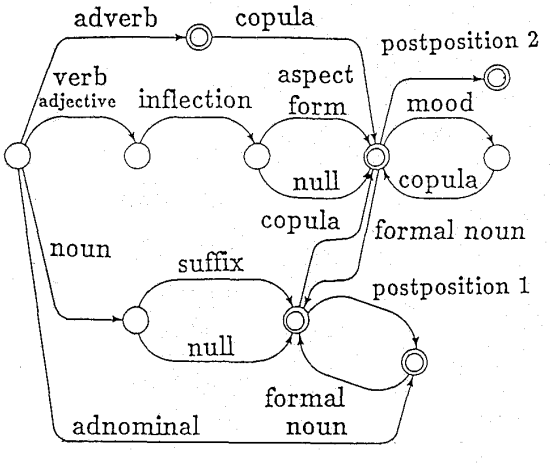

Figure 3 The outline of the structure of a Bunsetsu unit

VI. EXPERIMENTAL RESULTS

The training of the modified LVQ2 method was carried out for speech samples in the 212 word vocabulary uttered by 7 male and 8 female speakers. The recognition experiments of 30 phonemes were carried out for speech samples in the 212 word vocabulary uttered by another 3 male and 2 female speakers. Table 1 shows phoneme recognition scores. The result of N=2 corresponds to the original LVQ2 method. The recognition scores for $N \geq 3$ are higher than the score for N=2. This indicates the superiority of the modified LVQ2 method to the original LVQ2 method.

We applied this method to a multi-speaker dependent phoneme recognition task for continuous speech uttered Bunsetsu by Bunsetsu. Table 2 shows the phoneme recognition scores for 2 male speakers. The training for the modified LVQ2 method was carried out

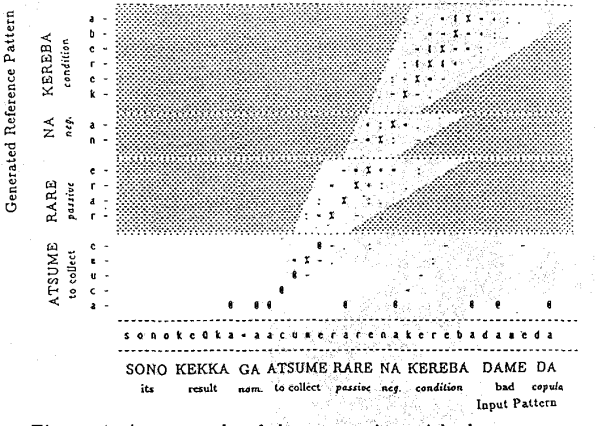

Figure 4 An example of the processing with the syntax driven continuous DP

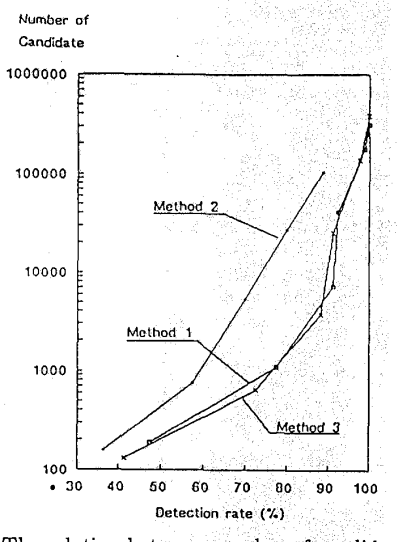

Figure 5 The relation between number of candidates per 100 input phonemes and the detection score when the phoneme recognition score is 85%.

using 70 sentences uttered by the two male speakers, where each of the two speakers uttered 35 sentences. The recognition experiments were carried out for the other 113 sentences uttered by each of the two speakers. The average phoneme recognition score was 86.1%. The average insertion and deletion scores were 7.7% and 3.9%. Table 3 shows recognition scores of the conceptual word, the Bunsetsu-unit and the sentence. The average recognition scores of the conceptual word, the Bunsetsu-unit and the sentence were 85.7%, 73.2% and 32.6%. Most sentence recognition errors are due to errors in recognition of functional words and in recognition of phonemes at the end of sentence.

VII. CONCLUSION

We have developed a prototype of a Japanese text dictation system which is composed of an acoustic processor, a Bunsetsu-unit spotting processor and a syntactic processor with semantic constraints. The acoustic processor is constructed using the modified LVQ2 method. The modified LVQ2 method achieves a high phoneme recognition performance of 86.1%. The syntax driven continuous DP matching algorithm is used for spotting Bunsetsu-units. This method greatly reduces the computation amount and storage capacity necessary for spotting the Bunsetsu-units. Analysis of dependency structure among the Bunsetsu-unit candidates is effectively carried out using the syntactic and semantic information.

REFERENCES

[1] M. Endo, S. Makino and K. Kido, "Phoneme recognition using a LVQ2 method", Trans. IEICEJ, SP89-50 (September, 1989) (in Japanese)

[2] Okada M., A. Ito, H.Matsuo, S.Makino and K. Kido, "Analysis of Japanese dictation system", Trans. Speech IEICEJ, SP86-33 (July, 1988) (in Japanese)

[3] Okada M., S.Makino and K. Kido, "A study of morphemic and syntactic processing sub-system for Japanese dictation system", Trans. IEICEJ, SP86-71 (December, 1986) (in Japanese)

[4] Ito A., Y.Ogawa, S. Makino and K.Kido, "Refinement and evaluation of Bunsetsu automaton in Japanese dictation system", Proc. ASJ meeting, pp. 135-136 (October, 1987) (in Japanese)

[5] Ogawa Y., A. Ito, M. Okada, S. Makino and K. Kido, "Refinement of syntactic processor in Japanese dictation system using semantic information", Proc. ASJ meeting, pp. 137-138 (October, 1987) (in Japanese)

[6] A. Ito, S. Makino and K. Kido, "A parsing algorithm based on CYK algorithm for continuous speech recognition", Proc. ASJ meeting, pp. 91-92 (October, 1988) (in Japanese)

[7] A. Ito, S. Makino and K. Kido, "Syntactic processing using the principle of least Bunsetsu's number method for continuous speech recognition", Proc. ASJ meeting, pp. 93-94 (October, 1988) (in Japanese)

[8] S.Moriai, S.Makino and K.Kido,"A method for selecting an optimum phoneme sequence using a posteriori probabilities of phonemes", Journal of ASA. supplement No. 1, PPP5 (November, 1988)

[9] T.Kohonen, G.Barna and R. Chrisley,"Statistical pattern recognition with neural networks: Benchmarking studies", IEEE Proc. of ICNN, Vol. 1, pp. 61-68 (July, 1988)

[10] E. McDermott and S. Katagiri,"Shift-invariant phoneme recognition using Kohonen networks", Proc. ASJ meeting, pp. 217-218 (October, 1988)

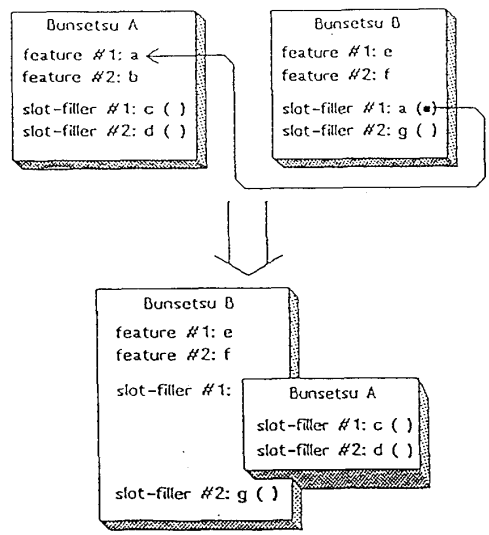

Figure 6 An example of the merging of two Bunsetsu units

Table 1 Speaker-independent phoneme recognition scores for spoken words using the modified LVQ2 method and the method for selecting an optimum phoneme sequence

Rank of reference vector for training	Phoneme recognition score	Deletion score	Insertion score
N=2	83.1	2.0	11.3
N=3	85.6	1.9	9.8
N=7	86.5	1.7	9.0

Table 2 Multi-speaker dependent phoneme recognition scores for continuous speech uttered Bunsetsu by Bunsetsu

Speaker	Phoneme recognition score	Deletion score	Insertion score
A	84.4	4.7	5.8
B	87.8	3.0	9.5

Table 3 Multi-speaker dependent Bunsetsu-unit recognition scores for continuous speech uttered Bunsetsu by Bunsetsu

Speaker	Conceptual word	Bunsetsu-unit	Sentence
A	84.8	70.9	28.4
B	86.7	75.6	36.7