論文 / 著書情報
Article / Book Information

Title	A Continuous Speech Recognition System Based on a Two-Level Grammar Approach
Author	Shoichi Matsunaga, Shigeki Sagayama, Shigeru Homma, Sadaoki Furui
Journal/Book name	IEEE ICASSP1990, Vol. , No. , pp. 589-592
発行日 / Issue date	1990, 4
権利情報 / Copyright	(c)1990 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Reprinted from PROCEEDINGS OF THE ICASSP 90 — 1990 INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Albuquerque, New Mexico, April 3-6, 1990

S11.7

A CONTINUOUS SPEECH RECOGNITION SYSTEM BASED ON A TWO-LEVEL GRAMMAR APPROACH

Shoichi Matsunaga, Shigeki Sagayama, Shigeru Homma and Sadaoki Furui

NTT Human Interface Laboratories
Midori-cho, Musashino-shi, Tokyo 180 Japan

ABSTRACT

This paper describes a Japanese continuous speech recognition system based on phonetic hidden Markov models (HMMs) combined with two levels of grammatical representations: an intra-phrase transition network grammar and an inter-phrase dependency grammar. A joint score, combining acoustic likelihood and linguistic certainty factor derived from phonetic HMMs and two levels of grammar, is maximized to obtain the optimal recognition results of sentences. Two efficient algorithms, bi-directional network parsing and breadth-first dependency parsing, are devised to globally optimize the joint score. The system attains a phrase recognition rate of 80.8% with the intra-phrase parser only, and 86.8% with both the intra-phrase and inter-phrase parsers, where the perplexity of the phrase syntax is 40. This result shows the effectiveness of the two-level grammar approach.

I. INTRODUCTION

Linguistic processing for speech recognition has been intensively studied from various approaches such as stochastic, syntactic and semantic grammars [1-3]. These grammars have advantages as well as limitations. Syntactic grammar is effective in describing the structure of phrases, while supposedly inadequate for describing the structure of whole sentences. On the other hand, case frame grammar [4] is suitable for governing the representation of the sentence structure. Thus, combining different types of grammars is practical for linguistic processing of speech recognition.

In Japanese sentences, which are sequences of minimal phrases, the phrase order is much less constrained than in English. On the other hand, the word order of phrases, which are short sequences of words, is very regular, and the sentence structure is ordered by semantic dependency between phrases. Syntactic constraints are useful in recognizing specific tasks or short-duration utterances. However, particularly in sentence recognition for phrase-order-free languages such as Japanese, semantic constraints are more powerful than sentence syntactic constraints.

We have developed a Japanese continuous speech recognition system which obtains the most likely sentence results taking account of acoustic, syntactic and semantic factors based on a two level grammar approach. This approach uses two grammars which are an intra-phrase transition network grammar for phrase recognition and an inter-phrase dependency grammar for sentence recognition. The former is a syntactic grammar and the latter is a semantic and loose syntactic grammar. The dependency grammar is compatible with the case grammar, and has robustness against missing or misrecognized words.

Two efficient parsing algorithms are devised for each grammar. They are a bi-directional network parser and a breadth-first dependency parser.

The syntactic structure within phrases is represented by recursive transition networks (RTNs) to concisely cover a variety of phrase structures. With the network parser, input phrase utterances are parsed bi-directionally both left-to-right and right-to-left to reduce the amount of computation, and optimal Viterbi paths are found along which the accumulated phonetic likelihood is maximized.

With the dependency parser, inter-phrase dependency structures within a sentence are analyzed. Semantic certainty factor is determined taking into account grammatical cases incorporated in word dictionaries. The joint score, obtained by combining accumulated phonetic likelihood and semantic certainty factor derived from the dependency grammar, is maximized to obtain the optimal solution. The dependency parser utilizes efficient breadth-first search and beam search algorithms.

The approach described here is highly suitable for speech understanding systems since it can use semantic dependency structures. Furthermore, it is applicable to a wide range of tasks since it does not need any sentence syntax.

II. SPEECH RECOGNITION SYSTEM USING TWO LEVEL GRAMMAR

A block diagram of the system is shown in Figure 1. Input sentences are uttered phrase by phrase. After feature parameter extraction of the utterance, the parameter sequence is converted into a vector code sequence. Next, phonetic likelihood is calculated for every possible duration to obtain likelihood matrices for phoneme candidates, based on HMMs. Phonetic duration time can be easily controlled by giving maximum and minimum duration times to each phoneme. Next, phrase likelihood is calculated based on the phonetic likelihood matrices. The verb and adjective entries in the dictionary have

CH2847-2/90/0000-0589 $1.00 © 1990 IEEE

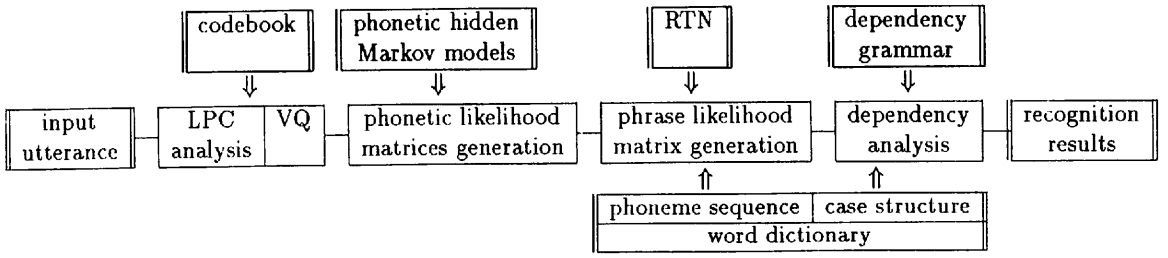


Figure 1: Block diagram of the continuous speech recognition system

grammatical cases, and the noun entries are accompanied by semantic primitives. Next, the top candidates of each likelihood are generated in a matrix form. The number of candidates is optional. Finally, using this matrix and the dependency grammar, the parser extracts the most likely sentence of a phrase sequence and its dependency structure.

III. INTRA-PHRASE SYNTACTIC GRAMMAR

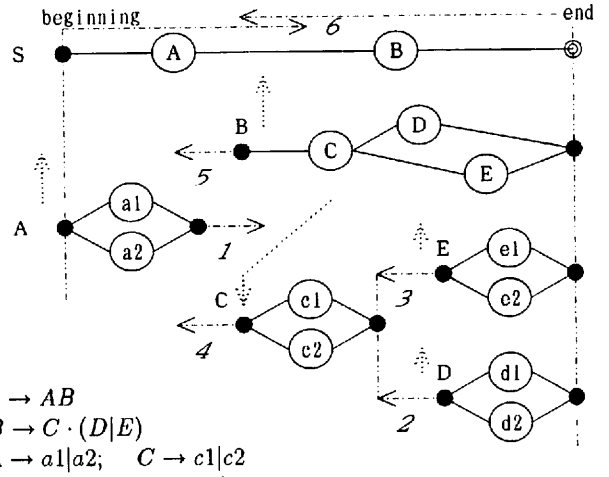
3.1 Syntactic Constraints of Phrases

Japanese phrases are composed of a stem part and a suffix part. The stem part is a verb, a noun, or an adjective. The suffix part is composed of suffixes such as auxiliary verbs and particles. Although connection of these words is very regular, there are many kinds of connection rules. Thus, to cover concisely this variety, the syntactic structure is represented by RTNs composed of sub-networks.

3.2 Speech Recognition of Phrases

When the phrase parser expands the RTNs to a single network in the phrase recognition process, the complexity of the phrase structure makes the network too large. Recognition is also time consuming. To overcome these problems, a bi-directional parser for RTNs is developed. This parser parses the input left-to-right in the stem part, and right-to-left in the suffix part. Bottom-up parsing is carried out for the stem words and the last suffixes of the suffix part, and top-down parsing is carried out for preceding suffixes. After calculating top-most m word candidates for each sub-network based on the Viterbi algorithm, the top-most m phrase candidates are generated.

A simple example of the phrase syntax and parsing flow is shown in Figure 2. S and B correspond to networks of the syntax, and A, C, D and E correspond to word sub-dictionaries. Processing flow is $A \rightarrow D \rightarrow E \rightarrow C \rightarrow B \rightarrow S$, and the processing order number is shown in the figure. First, the stem words, corresponding to A, are processed left-to-right from the beginning of speech. Next, the last suffixes, corresponding to D and E, are processed right-to-left from the end of speech. The calculation result of B is generated by combining the results of D, E and C. The likelihood calculation of C, namely the likelihoods of $c1$ and $c2$, is carried



$S \rightarrow AB$
$B \rightarrow C \cdot (D|E)$
$A \rightarrow a1|a2; \quad C \rightarrow c1|c2$
$D \rightarrow d1|d2; \quad E \rightarrow e1|e2$

Figure 2: Example of phrase recognition process

out at every top-down scan of C. Finally, the phrase recognition results are generated by combining the results of A and B. In the figure, arrows $\rightarrow$, $\leftarrow$, $\uparrow$, $\downarrow$ indicate the left-to-right process, the right-to-left process, the bottom-up control and the top-down control, respectively.

IV. INTER-PHRASE DEPENDENCY GRAMMAR

4.1 Semantic and Loose Syntactic Constraints

Dependency grammar is based on semantic dependency relationships between phrases. The syntactic rules satisfy the only two constraints. First, every phrase except the last must modify one and only one later phrase. This modification is called a dependency relationship or dependency structure. Second, no modification relationship between phrases in the sentence cross.

The semantic certainty factors of the dependency structure are easily provided using grammatical cases. There are two kinds of factors. One is associated only with dependency relationships of the modifier and modificant phrases: agreement between the semantic primitive of the modifier and that required by the modificant, agreement between the case of modifier and that required by the modificant, idiomatic expressions and so on. The other factor is associated with all

the dependency structures of the phrase sequence: a phrase with the obligatory case required by the modificant, no modification of the same phrase by different phrases having the same case, simplicity of the sentence structure and so on. The certainty factor values for these items are given heuristically.

4.2 Parser for Dependency Structure Grammar

This parsing is equivalent to solving the following objective function using the constraints of dependency structure grammar.

$$T = \max_{\{p\}}[\sum_{j=1}^{N} c(x_{j,p}) + \max_{Y} \sum_{j=1}^{N} dep(\boldsymbol{w}_{1,j-1}, x_{j,p}|Y_{1,j,p})] \quad (1)$$

where $1 \leq j \leq N$, $1 \leq p \leq M$, N is the number of input phrases, M is the maximum number of recognition candidates for each uttered phrase, $x_{j,p}$ is a candidate of the j-th input phrase with the p-th best likelihood, and $c(x_{j,p})$ is its log-likelihood. A phrase sequence with one phrase candidate for each i-th to j-th input phrase and whose last phrase is $x_{j,p}$ is denoted by $X_{i,j,p}$. $Y_{i,j,p}$ is one of the dependency structures of $X_{i,j,p}$, $\boldsymbol{w}_{i,j-1}$ is the set of phrases that modify $x_{j,p}$ in the sequence $X_{i,j,p}$. Here, $dep(\boldsymbol{w}, x|Y)$ is the linguistic certainty factor of dependency relationships between $\boldsymbol{w}$ and x taking Y into account. The first item of the term on the right in Eq.(1) is the summation of phonetic likelihoods of the hypothesized sentence composed of its phrase sequence, and the second item is the summation of linguistic certainty factor. Maximizing Eq.(1) gives the sentence and its dependency structure as the speech recognition result.

To solve Eq.(1) effectively, a fast parsing algorithm using breadth-first search and beam search was developed. This algorithm is based on the fundamental algorithms [5,6]. Although it offers sub-optimal solutions, it is practical because it requires much less processing than the depth-first search.

The breadth-first algorithm is formulated as follows. Its derivation is described in detail in reference [7]. First, $dep(\boldsymbol{w}, x|Y)$ can be divided into two terms.

$$dep(\boldsymbol{w}_{1,j-1}, x_{j,p}|Y_{1,j,p}) = \sum_{x \in w_{1,j-1}} dep1(x, x_{j,p}) + dep2(Y_{1,j,p}, x_{j,p}) \quad (2)$$

where $dep1$ is the certainty factor associated with dependency relationships of only the modifier and modificant phrases, and $dep2$ is the certainty factor associated with $Y_{1,j,p}$. Using notation $S(1, x_{j,p})$, the objective function's value of a phrase sequence including the top phrase to $x_{j,p}$ in the sentence, and $D(i, x_{j,p})$, the value of a phrase sequence not including the top phrase ($i \neq 1$), the recursive relation using beam search are derived.

$$S(1, x_{j,p}, r) = r^{th}\max_{k,q,r1,r2}[S(1, x_{k,q}, r1) + D(k+1, x_{j,p}, r2) + dep1(x_{k,q}, x_{j,p}) + dep2(Y_{1,j,p}, x_{j,p})] \quad (3)$$

$$D(i, x_{j,p}, r) = r^{th}\max_{k,q,r1,r2}[S(i, x_{k,q}, r1) + D(k+1, x_{j,p}, r2) + dep1(x_{k,q}, x_{j,p}) + dep2(Y_{i,k,q}, x_{k,q})] \quad (i \neq 1) \quad (4)$$

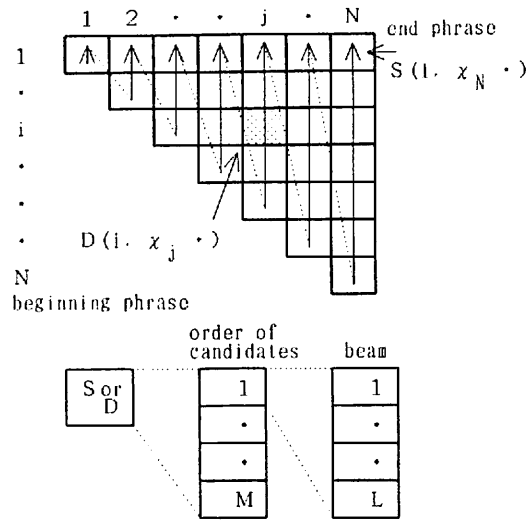


Figure 3: Configuration of parsing table

where $i \leq k \leq j-1$, $1 \leq q \leq M$,and $1 \leq r, r1, r2 \leq L$. Here, $r, r1$ and $r2$ indicate the beam ranks, L is the maximum number of beams, $S(1, x_{j,p}, r)$ and $D(i, x_{j,p}, r)$ are the r-th value of the element whose phrase sequence is $X_{i,j,p}$ and the dependency structure is $Y_{i,j,p}$. Here, $r^{th}\max[\cdot]$ is a function for deriving the r-th best value. When Eq.(3) or (4) is calculated, $Y_{i,j,p}$ is stored for use in the later stage of evaluating $dep2$.

Initial values are given as follows.

if $i = 1$ (top phrase)

$$S(1, x_{1,p}, 1) = c(x_{1,p}) + dep2(Y_{1,1,p}, x_{1,p}) \quad (5)$$

if $i \neq 1$ (not top phrase)

$$D(i, x_{i,p}, 1) = c(x_{i,p}) \quad (6)$$

After calculating the recurrence relation, the value of the objective function is obtained:

$$T = \max_{p}[S(1, x_{N,p}, 1)] \quad (7)$$

where $1 \leq p \leq M$. The best sentence and its dependency structure are given through $Y_{1,N,p}$ where p maximizes Eq.(7). The parsing table is shown in Figure 3 and the parsing algorithm is shown in Table 1. In Figure 3, the first row corresponds to S, and others correspond to D. The phrase sequence for the first to N-th phrase corresponds to the rightmost top cell. Each cell is composed of M sub-cells for the number of candidates, and each sub-cell is composed of L sub-cells for the beam width. The arrows show the sequence of calculating the recurrence relation. The processing amount order for this algorithm is $O(N^3M^2L^2)$.

V. SPEECH RECOGNITION EXPERIMENTS

An input utterance was sampled at a rate of 12 kHz. One frame was extracted every 10ms with a 30ms Hamming

Table 1: Parsing algorithm

{1} : Loop for the end of phrase of the partial sequence
DO {2} to {5} for $j = 1, 2, \ldots, N$
{2} : Loop for the candidates
DO {3} to {5} for $p = 1, 2, \ldots, M_j$
{3} : Setting the initial value
SET $S(1, x_{1,p}, 1)$ or $D(j, x_{j,p}, 1)$ (Eqs.(5),(6))
if $j = 1$, go back to {2}.
{4} : Loop for the beginning phrase of the partial sequence
DO {5} for $i = j-1, j-2, \ldots, 1$
{5} : Computation of recurrence relation
(*Loop for the end phrase of the former sequence*)
{5 − 1} : DO {5-2} to {5-4} for $k = i, i+1, \ldots, j-1$
{5 − 2} : DO {5-3} to {5-4} for $q = 1, 2, \ldots, M_k$
(*Loop for the beam width*)
{5 − 3} : DO {5-4} for $r1 = 1, 2, \ldots, L$
{5 − 4} : for $r2 = 1, 2, \ldots, L$
* Evaluation of $S(1, x_{1,p}, r)$ or $D(j, x_{j,p}, r)$ taking account of $Y_{1,j,p}$ or $Y_{i,k,q}$ (Eqs.(3),(4))
* Store of $Y_{i,j,p}$
{6} : Acquisition of the parsing results
* Detection of value p maximizing Eq.(7)
* Acquisition of the phrase sequence and its dependency structure using $Y_{1,N,p}$

window and converted into 34 acoustic feature parameters: power, 16 LPC cepstra, Δ power and 16 Δ LPC cepstra [8,9].

In the training process, 216 phonetically balanced words were used. These utterances were manually labeled using 25 phoneme symbols including silence. Each phoneme was modeled by HMM and had 4 states and 7 transition paths. The parameter sequence was converted into a vector code sequence and a vector codebook composed of 256 prototype vectors was generated. Training of each HMM was carried out using the forward-backward algorithm, and the code sequence for training was cut out based on the phonetic labels. Output probabilities were floored after training.

In the testing process, input sentences were uttered phrase by phrase. Acoustic feature parameters of the input were generated in the same manner. These parameters were converted into a code sequence using the same speaker's codebook generated in the training process.

First, talker-dependent preliminary recognition tests were done on two sets of 216 words, one for training and the other for test, uttered by 10 speakers, 5 males and 5 females. The system attained a word recognition rate of 99.9% for the training set and 98.4% for the test set.

Next, talker-dependent recognition tests were performed on 100 sentences (including 668 phrases in an essay) uttered by the same speakers. The word dictionary had 360 entries and the perplexity of the phrase syntax was 40. The phrase syntax was capable of generating about 10^4 hypotheses for each phrase of speech. The number of candidates, M, was 5, and the beam width, L, was 8. Certainty factor values were empirically determined through feature-based speech recognition [7] of technical literature.

The system attained a phrase recognition rate of 80.8% using the intra-phrase syntactic parser only. The dependency parser increased this rate to 86.8%, as shown in Table 2. This result shows the effectiveness of the two-level grammar approach.

Table 2: Speech recognition results

data	216 words		668 phrases	
	training	test		
dependency analysis	–		without	with
recognition rate	99.9	98.4	80.8	86.8
(top 5)	100	100	96.8	–

VI. CONCLUSION

This paper described a Japanese continuous speech recognition system using an intra-phrase transition network grammar and an inter-phrase dependency grammar. Input utterances were recognized efficiently to determine the best sentence using a bi-directional network parser and a breadth-first dependency parser. Recognition experiment results showed the effectiveness of the inter-phrase dependency grammar. The parser for this grammar can be easily expanded for sentence speech recognition [7].

Further development is currently in progress to refine phonetic models based on continuous HMMs which take context dependency into account.

ACKNOWLEDGMENT

The authors wish to express their appreciation to Hirokazu Sato of NTT Labs and Masaki Kohda of the University of Yamagata for their invaluable discussions. The authors also thank Frank Soong and Fred Juang of AT&T Bell labs for their useful suggestions.

REFERENCES

[1] V.R.Leser, et al "Organization of the Hearsay II speech understanding system", *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-23 No 1, pp.11-24, 1975
[2] H.Ney, "Dynamic programming speech recognition using a context-free grammar", *Proc. ICASSP*, pp.69-72 , 1987, Dallas
[3] K-F.Lee and H-W. Hon, "Large-vocabulary speaker-independent continuous speech recognition using HMM", *Proc. ICASSP*, pp.123-126 , 1988, New York
[4] C.Filmore, "The case for case." in Bach and Harms(eds.), 1-88, 1968
[5] S.Matsunaga and M.Kohda, "Post-processing using dependency structure of inter-phrases for speech recognition." 1-1-23, *Proc. ASJ annual meeting*, pp.45-46, Mar.1986 (in Japanese)
[6] K.Ozeki,"A multi stage decision algorithm for optimum bunsetsu sequence selection." *Paper Tec. Group*, IECE Japan, SP86-32, pp.41-48, Jul.1986 (in Japanese)
[7] S.Matsunaga and M.Kohda,"Linguistic processing using a dependency structure grammar for speech recognition and understanding.", *Proc. COLING*, pp.402-407, 1988, Budapest
[8] S.Furui,"Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34 No 1, pp.52-59, 1986
[9] S.Sagayama and F.Itakura, "On individuality in a dynamic measure of speech." 3-2-7, *Proc. ASJ annual meeting*, pp.589-590, Jul.1979 (in Japanese)