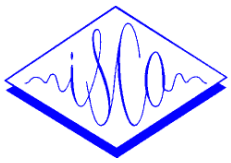

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

CONTINUOUS SPEECH RECOGNITION USING TWO-LEVEL LR PARSING

Kenji Kita, Toshiyuki Takezawa, Junko Hosaka, Terumasa Ehara, Tsuyoshi Morimoto

ATR Interpreting Telephony Research Laboratories
Seika-cho, Souraku-gun, Kyoto 619-02, Japan

ABSTRACT

This paper describes a continuous speech recognition system using two-level predictive LR parsing. ATR has already implemented a predictive LR parsing algorithm in an HMM-based speech recognition system for Japanese. However, up to now, this system has used only intra-phrase grammatical constraints. In Japanese, a sentence is composed of several phrases, thus two kinds of grammars, namely an intra-phrase grammar and an inter-phrase grammar, are sufficient for recognizing sentences. Two-level predictive LR parsing makes it possible to use not only intra-phrase grammatical constraints but also inter-phrase grammatical constraints during speech recognition. The system is applied to Japanese sentence recognition, and attains a word accuracy of 95.9% and a sentence accuracy of 84.7%.

1 INTRODUCTION

Speech recognition alone will not attain good performance for large vocabulary tasks. To improve recognition accuracy, some kinds of linguistic information are necessary, such as syntax, semantics, pragmatics, and dialogue. Currently, most speech recognition systems use syntax which is more tractable than other linguistic information.

In Japanese, a sentence is composed of several phrases, thus using two kinds of grammars, namely an intra-phrase grammar and an inter-phrase grammar, is sufficient for recognizing sentences. For example, Matsunaga et al. [1] use an intra-phrase transition network grammar and an inter-phrase dependency grammar to recognize Japanese sentences.

ATR has already implemented a predictive LR parsing algorithm in an HMM-based speech recognition system, and successfully applied it to Japanese phrase recognition [2, 3]. Basically, this system, called *HMM-LR*, can deal with grammatical constraints based on a context-free grammar. However, up to now, this system uses only intra-phrase grammatical constraints. That is, inter-phrase grammatical constraints are not considered during speech recognition. The system often outputs ungrammatical recognition results as sentences. For using HMM-LR in spoken-language processing applications such as a voice translation system, it is quite desirable to use inter-phrase grammatical constraints which can capture the global syntax of a language.

We developed a two-level LR parsing that can easily deal with both intra-phrase grammatical constraints and inter-phrase grammatical constraints. The two-level LR parsing is applied to Japanese sentence recognition where sentences were uttered phrase by phrase. Currently, the system attains a word accuracy of 95.9% and a sentence accuracy of 84.7%.

2 HMM-LR SPEECH RECOGNITION SYSTEM

First, we will review the HMM-LR speech recognition system [2]. The system consists of a *predictive LR parser* and *HMM phone verifiers*. The predictive LR parser is based on the *generalized LR parsing* [4], and is used not only for parsing but also phone prediction. The parser is guided by an LR parsing table which is pre-compiled from context-free grammar rules.

The actual recognition process is as follows. First, the system picks up all phones predicted by the initial state of the LR parsing table. Next, the system invokes the HMM phone models of these predicted phones to verify their existence. The system then proceeds to the next state in the LR parsing table. During this process, all possible partial parses are constructed in parallel. The HMM phone verifier receives a probability array which includes end point candidates and their probabilities, and updates it using the forward algorithm. This probability array is attached to each partial parse. The beam-search technique is also used, and partial parses with low likelihood score are pruned.

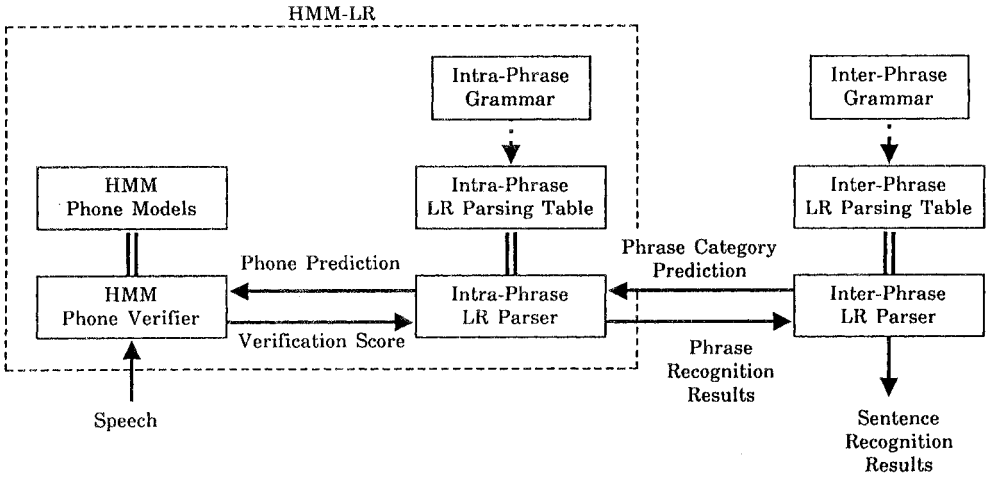

Figure 1: Continuous Speech Recognition System Using Two-Level LR Parsing

High recognition performance is attained by driving HMMs directly without any intervening structures such as a phone lattice. The part of Figure 1 enclosed with a dotted line shows the HMM-LR speech recognition system.

3 TWO-LEVEL LR PARSING

3.1 Predictive LR Parsing

In our approach, the LR parser can be regarded as a language source model for symbol prediction/generation. For example, in the HMM-LR, since its terminal symbols are phones, the predictive LR parser predicts next phones at each generation step and generates many possible sentences as phone sequences. The predictive LR parser determines next phones using the action table in the LR parsing table and splits the parsing stack not only for grammatical ambiguity but also for symbol variation.

The action table is a two-dimensional table, where one dimension is indicated by states and the other by terminal symbols. Thus if the state is given, terminal symbols which might come next will be obtained. To be more precise, terminal symbols associated with *shift* actions are predicted.

3.2 Two-Level LR Parsing

A two-level predictive LR parsing uses two kinds of grammars, an intra-phrase grammar and an inter-phrase grammar. Both grammars are written in the form of context-free grammar. These grammars are automatically compiled into an intra-phrase LR parsing table and an inter-phrase LR parsing table, respectively.

Unlike a standard LR parsing table, an intra-phrase LR parsing table consists of three parts: an *action table*, a *goto table* and a *goto-phrase table*. The third table is a new one, and decides to which state the parser should go when a phrasal category is given. That is, a goto-phrase table indicates the initial state of the parser for each phrasal category. An inter-phrase LR parsing table is the same as a standard LR parsing table.

Figure 1 shows a continuous speech recognition system using the two-level predictive LR parsing.

The recognition process is summarized below. First, the inter-phrase LR parser predicts next phrasal categories, e.g. NP (noun phrase), using the inter-phrase LR parsing table. Then, the intra-phrase LR parser decides the NP initial state using the goto-phrase table in the intra-phrase LR parsing table. The intra-phrase LR parser picks up all phones predicted by the NP initial state, and invokes the HMM phone models to verify the existence of these predicted phones. After recognizing NP candidates, the next phrasal category, e.g. VP (verb phrase), is predicted using the inter-phrase LR parsing table. The recognition process proceeds in this way until the entire speech data input is processed.

3.3 Construction of Intra-Phrase LR Parsing Table

This subsection describes how to construct an intra-phrase LR parsing table from the grammar. Here we use *Grammar 1* in Figure 2. In this grammar, the start symbol is indicated by $START$, and phrasal categories are NP and VP.

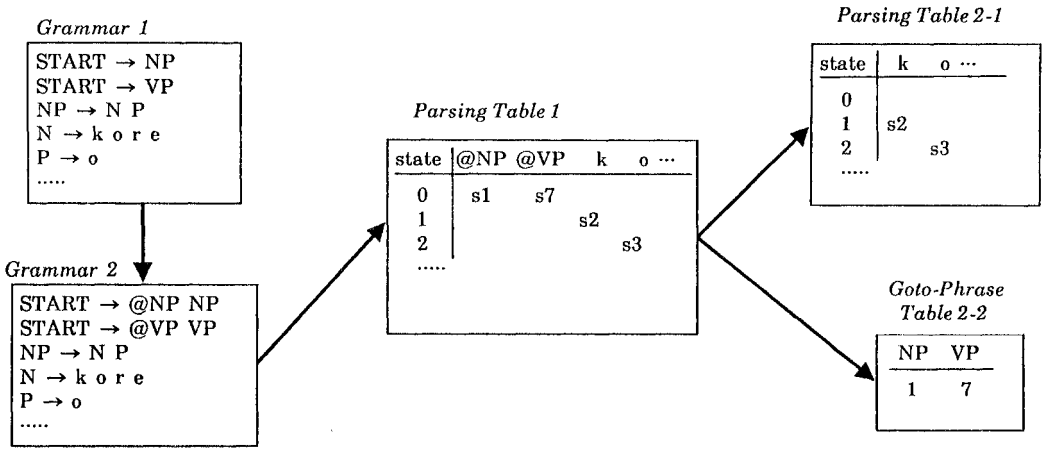

Figure 2: Construction of Intra-Phrase LR Parsing Table

1. For a rewriting rule $S \rightarrow A$ where S is a start symbol and A is a phrasal category, prepare a new terminal symbol $@A$ that does not appear in the grammar, and replace the rewriting rule with $S \rightarrow @AA$. In case of the *Grammar 1*, new terminal symbols are $@NP$ and $@VP$, and we get *Grammar 2*.
2. Construct *Parsing Table 1* from *Grammar 2* using a standard LR parsing table construction algorithm.
3. Remove the new terminal symbols from *Parsing Table 2*. We get *Parsing Table 2-1*.
4. Create *Goto-Phrase Table 2-2* by checking states after new terminal symbols are shifted at the initial state.

4 RECOGNITION EXPERIMENTS

4.1 HMM Phone Models

The HMMs used in the experiments are basically the same as those reported in [3].

HMM phone models are based on discrete HMM. A three-loop model for consonants and a one-loop model for vowels are trained using each phone data extracted from the ATR isolated word database [5].

Duration control techniques and separate vector quantization are used to achieve accurate phone recognition.

4.2 Speech Data

The recognition experiments were carried out using 137 sentences uttered by one male speaker. These sentences were uttered phrase by phrase. Sentence examples are shown in Table 1.

The speech data is sampled at 12kHz, pre-emphasized by $(1 - 0.97z^{-1})$, and windowed using a 256-point Hamming window every 9 msec. Then, 12-order LPC analysis is carried out. Spectrum (WLR), difference cepstrum coefficients, and power are computed.

Table 1: Sentence Examples

moshimoshi
(hello)
kaigini moushikomitainodesuga
(I would like to register the conference)
tourokuyoushiwa sudeni omochideshouka
(Do you already have a registration form?)

Table 2: Grammar Size and Complexity

	Intra-Phrase Grammar	*Inter-Phrase Grammar*
Rules	1,973	471
Vocabulary	744 words	133 phrasal categories
Perplexity	3.57 / phone	99.7 / phrase

4.3 Grammar

Two kinds of context-free grammars were used in the experimemts: an intra-phrase grammar and an inter-phrase grammar. These grammars describe the task of an *International Conference Secretary Service*. The phonetic lexicon for the task is embedded in the intra-phrase

Table 3: Recognition Rates

rank	*Phrase Lattice Parsing*		*Two-Level LR Parsing*	
	Word Accuracy	Sentence Accuracy	Word Accuracy	Sentence Accuracy
1	87.8	78.8	95.9	84.7
2	-	85.4	-	90.5
3	-	86.9	-	93.4

grammar. Grammar size and complexity are shown in Table 2.

4.4 Results

Two kinds of experiments were carried out. In the first experiment, intra- and inter-phrase LR parsers were used independently. That is, the intra-phrase LR parser was used to produce the top five phrase candidates. Then, the inter-phrase LR parser was applied to obtain sentence candidates. In the second experiment, two-level LR parsing was used.

Results are evaluated by *accuracy*, which can be calculated as follows.

$$accuracy = \frac{total - sub - ins - del}{total} \times 100$$

where *total* indicates the number of words in all the sentences, and *sub*, *ins* and *del* are the number of words recognized as the incorrect word, deleted and inserted, respectively. Results of recognition experiments are shown in Table 3. By using two-level LR parsing, word accuracy and sentence accuracy are improved from 87.8% and 78.8% to 95.9% and 84.7%, respectively.

In the first experiment, correct phrases were sometimes missed within top five choices. However, using inter-phrase grammatical constraints during recognition, the number of missed phrases decreased. This is why word accuracy and sentence accuracy are improved by using the two-level LR parsing.

5 CONCLUSION

This paper described a continuous speech recognition system using two-level predictive LR parsing. Two-level LR parsing makes it possible to use both intra- and inter-phrase grammatical constraints during speech recognition. The system attained word accuracy of 95.9% and sentence accuracy of 84.7%. These results show that two-level LR parsing is effective in sentence recognition.

Future works include using higher level linguistic information such as semantics, extending vocabulary size, and unknown word processing.

ACKNOWLEDGMENTS

The authors are deeply grateful to Dr. Kurematsu, the president of ATR Interpreting Telephony Research Laboratories, all the members of the *Speech Processing Department* and the *Knowledge and Data Base Department* for their constant help and encouragement.

REFERENCES

[1] S. Matsunaga, S. Sagayama, S. Homma and S. Furui, "A Continuous Speech Recognition System Based on a Two-Level Grammar Approach", ICASSP 90 (1990).

[2] K. Kita, T. Kawabata and H. Saito, "HMM Continuous Speech Recognition Using Predictive LR Parsing", ICASSP 89 (1989).

[3] T. Hanazawa, K. Kita, S. Nakamura, T. Kawabata and K. Shikano, "ATR HMM-LR Continuous Speech Recognition System", ICASSP 90 (1990).

[4] M. Tomita, "Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems", Kluwer Academic Publishers (1986).

[5] H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri and T. Watanabe, "Construction of a Large-Scale Japanese Speech Database and its Manegement System", ICASSP 89 (1989).