統計的言語情報を用いたHMM-LR文章発声音声認識の評価

北 研二　森元 逞　大倉 計美　嵯峨山 茂樹

ATR自動翻訳電話研究所

概要：　HMM-LR音声認識システムを連続発声の文認識に適用した結果について述べる。これまでに我々は、HMM-LR音声認識システムを日本語の文節発声の音声認識に適用し、高い認識率が得られることを示した。今回、文レベルでの連続音声認識を行なうにあたり、いくつかの改良を行なった。まず最初に、従来の文節認識実験では、HMM音韻モデルの学習データとして単語発声のデータのみを用いていたが、連続発声中での調音結合の影響に対処するために、連続発声のデータを追加した。次に、日本語の文に対する文法を作成することにより、文認識を可能とした。このため、文節内の文法的制約を記述した文節内文法と、文節間の文法的制約を記述した文節間文法をそれぞれ別個に開発し、これらの2つの文法を統合化することによって、文の文法を作った。更に、言語の統計的な情報を用いることにより、高精度の音声認識を目指した。ここでは、統計的な言語モデルとして、確率文脈自由文法と生成規則間のマルコフモデルを用いた場合について評価を行なった。以上の改良をほどこしたHMM-LR音声認識システムを、約750単語を含む「国際会議への参加登録申し込み」タスクで評価した結果、特定話者で83.9%の文認識率を達成した。

Evaluation of the HMM-LR Speech Recognition System against Continuous Sentential Utterances with the Aid of Stochastic Linguistic Knowledge

Kenji Kita, Tsuyoshi Morimoto, Kazumi Ohkura and Shigeki Sagayama

ATR Interpreting Telephony Research Laboratories

Abstract : This paper describes recent efforts to improve the HMM-LR speech recognition system for continuously spoken sentences. The HMM-LR system has been applied to Japanese phrase recognition and has attained high recognition performance. However, up to now, the system has not been applied to continuous spoken sentence recognition. In this work, several improvements have been made on the system. The first improvement is HMM training with continuous utterances as well as word utterances. In previous implementation, HMMs have been trained with only word utterances. Continuous utterances are included in HMM training data because coarticulation effects are much stronger in continuous utterances. The second improvement is the development of a sentential grammar for Japanese. The sentential grammar was created by combining inter- and intra-phrase grammars, which were developed separately. The third improvement is the incorporation of stochastic linguistic knowledge, which includes stochastic CFG and an N-gram model of production rules. The system was evaluated using continuously spoken sentences from a conference registration task that includes approximately 750 words. A sentence accuracy of 83.9% was attained in the speaker-dependent condition.

1 INTRODUCTION

Speech recognition systems that take full advantage of context-free grammar (CFG) constraints are increasingly common. Also, the LR parsing technique has been extensively used in dealing with CFG constraints, due to its no-backtracking table-driven efficiency [1, 2, 3, 4, 5, 6]. The HMM-LR speech recognition system, which is an integration of phone-based HMMs and LR parsing, has provided high recognition performance for Japanese phrase recognition with the introduction of multiple codebooks, accurate HMM state duration control, and fuzzy vector quantization [7]. However, up to now, the system has not been applied to continuous utterances.

This paper describes recent improvements in the HMM-LR speech recognition system aimed at handling continuously spoken sentences. The following outlines the major improvements:

1. In the system currently used for phrase utterances, HMM phone models have been trained with isolated word utterances. However, phones in continuous utterances exhibit great variability due to strong coarticulation. To overcome this problem, HMMs were trained with continuous utterances as well as word utterances.

2. The grammatical structure of Japanese has two levels: the intra-phrase level and the inter-phrase level. Thus, intra- and inter-phrase CFG grammars are developed separately, and then combined into one sentential grammar. The sentential grammar allows pauses between phrases. The system recognizes naturally spoken sentences including phrase utterances and continuous utterances.

3. In order to account for the stochastic aspects of a language, two kinds of stochastic knowledge are investigated: a stochastic CFG and an N-gram model of production rules. The latter is a particularly powerful stochastic language model that takes context-sensitivity into account.

This paper first gives an overview of the HMM-LR speech recognition system. Then recent improvements and experiments evaluating performance are described.

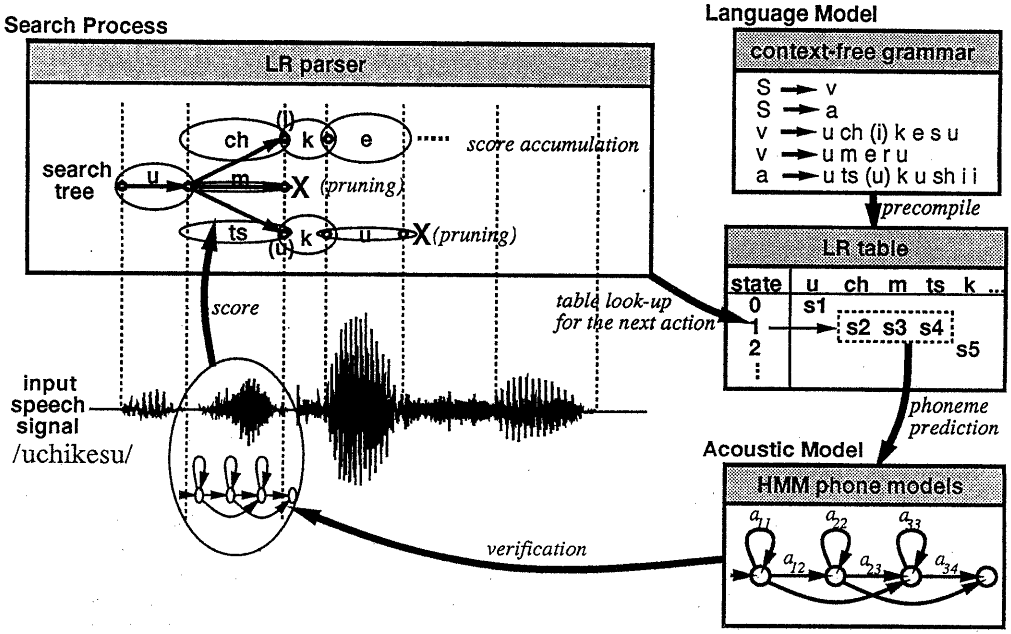

Figure 1: HMM-LR speech recognition system.

2 HMM-LR SPEECH RECOGNITION SYSTEM

The HMM-LR speech recognition system is depicted in Figure 1. The system uses phone-based HMMs for acoustic modeling and an LR parser for dealing with grammatical constraints based on a context-free grammar.

2.1 Acoustic Component

Phone-based HMMs are used for acoustic modeling. HMM phone models are based on discrete HMMs. Vowels, the syllabic nasal and a silence are modeled as a one-loop model, and other phones as a three-loop model.

For accurate phone modeling, separate vector quantization (multiple codebooks) is used, where the following three parameters are vector-quantized separately:

- Spectrum (WLR)
- LPC cepstral difference
- Power

HMM state duration control is also introduced. State duration is approximated by the Gaussian distribution through the process of Viterbi alignment for training data, and is used as a penalty for transition and output probabilities.

2.2 Linguistic Component

The linguistic component of the HMM-LR system is based on *predictive LR parsing* [3]. Predictive LR parsing is an extension of *generalized LR parsing* [8], which has symbol prediction capability. In other words, it predicts possible subsequent symbols by refering to a parsing table. This prediction process is based on a table look-up, and can thus be done very efficiently. Accordingly, predictive LR parsing provides a computationally inexpensive and powerful mechanism for search space reduction.

We use the predictive LR parser for guiding the search of an HMM-based speech recognizer. During recognition, the parser predicts at each stage of the parse all possible subsequent phone hypotheses. Each hypothesis is then assigned a probabilistic likelihood value from its corresponding HMM. Recognition hypotheses with low likelihood values are pruned by a beam-search technique. This integration of the predictive LR parser and phone-based HMMs allows a tight coupling of speech recognition and natural language processing and results in a computationally efficient algorithm.

3 SPEECH DATA

It is well known that HMMs perform better with more training data. It is also well known that phones in continuous utterances exhibit great variability due to strong coarticulation. The previous attempt at phrase recognition used HMMs trained by word utterances, which included 5,240 important Japanese words and 216 phonetically balanced words.

In this work, phrase and continuous utterances were extracted from the *ATR speech database* [9] and included in the training data. Table 1 shows speech data used for HMM training or evaluation. All speech data were uttered by one male speaker, recorded in noise-free environments, and phonetically labeled by hand.

Table 1: Training and evaluation speech data.

Speech Data	*Utterance Type*	*Number*
Training Data	Word	5,240 important Japanese words & 216 phonetically balanced words
	Phrase	598 phrases
	Continuous	90 sentences
Evaluation Data	Continuous	137 sentences

The speech data is sampled at 12 KHz, pre-emphasized with a filter having a transform function of $(1-0.97z^{-1})$, and windowed using a 256-point Hamming window every 9 msec. Then, 12-order LPC analysis is carried out, and finally the VQ code sequence is generated. For VQ codebook generation, 216 phonetically balanced words are used.

4 SENTENTIAL GRAMMAR

Japanese is very different from European languages such as English. The grammatical structure of Japanese has two levels: the intra-phrase level and the inter-phrase level. The intra-phrase level constraints are well described by a regular grammar or a CFG. As for the inter-phrase level constraints, most Japanese speech recognition systems use a semantic relationship called the *kakariuke relationship*. However, it is very difficult to make kakariuke relationships for large vocabulary tasks. Moreover, kakariuke relationships are very difficult to utilize for dynamically constraining the search space during recognition.

In our implementation, both intra- and inter-phrase level constraints are described in CFG form. First, the intra- and inter-phrase CFGs were developed separately [10], and then these CFGs were combined into one sentential grammar. Terminal symbols in the intra- and inter-phrase CFGs are phone names and phrase category names, respectively. Table 2 shows the size and complexity of each grammar. The test-set perplexity of the sentential grammar is 56/word.

In Japnese, since a phrase is both a grammatical and a phonological unit, it is uttered in one breath in a natural speaking style. Thus, the sentential grammar allows pauses between phrases.

Table 2: Size and complexity of the intra- and inter-phrase grammars.

Grammar	*Intra-Phrase*	*Inter-Phrase*
Rules	1,973 rules	471 rules
Vocabulary	744 words	133 phrase categories
Perplexity	3.57/phone	65.5/phrase

5 STOCHASTIC LANGUAGE MODELS

A pure CFG model inherently contains the following problems:

- **Overgeneration problem**
  A CFG is created by hand-parsing a corpus of text based on human linguistic knowledge. However, the resulting CFG will not only generate correct sentences, but also many other incorrect sentences. We carried out a sentence generation experiment to see how many incorrect sentences were generated from our sentential grammar. As a result, about 80% of all generated sentences were incorrect [11].

- **Ambiguity problem**
  The number of syntactic ambiguities in one sentence becomes increasingly unmanageable as the number of phrases increases. Martin et al. [12] reported that the syntactic ambiguities of sentences can be approximated by the Catalan number, which grows almost exponentially.

To address these problems, it is advantageous to use stochastic language models. A stochastic language model assigns a high probability to a correct sentence and a low probability to an incorrect sentence. It also provides an effective means for sentence disambiguation.

To combine the advantages of syntactic and stochastic language models, the following language models are investigated:

- Stochastic CFG

- N-gram model of production rules (rule N-gram)

These models provide hybrid modeling of a language, in which stochastic models augment a syntactic model quantitatively.

A Stochastic CFG assumes a strong assumption, namely a probabilistic independent assumption, that the choice of the production rule is independent of the context. Recently, more powerful language models beyond stochastic CFGs have attracted considerable attention [13, 14, 15, 16], where some models take context-sensitive probabilities into account. An N-gram model of production rules is such an attempt. Actually, we use the model where N equals 2, which is the bigram model of production rules.

5.1 Stochastic CFG

A stochastic CFG [17] extends a CFG so that each production rule is of the form $<A \rightarrow \alpha, p>$, where p is the conditional probability of A being rewritten into α. The probabilities of all A-productions (rules having A on the LHS) should sum to 1.

In the stochastic CFG, the probability of a derivation can be computed as the product of the probabilities of the rules used. Suppose that

$$S \xRightarrow{r_1} \gamma_1 \xRightarrow{r_2} \gamma_2 \xRightarrow{r_3} \cdots \xRightarrow{r_n} \gamma_n = x \tag{1}$$

is a derivation of x from the start symbol S, then the probability of this derivation is given by

$$P(x) = \prod_{i=1}^{n} P(r_i). \tag{2}$$

A stochastic CFG can be converted into a stochastic LR parsing table, in which each action entry contains a stochastic factor [4]. The LR parser uses these stochastic factors in the computation of the runtime stochastic product, which reflects the likelihood of each recognition hypothesis.

5.2 N-gram Model of Production Rules

The structure of a sentence is represented as its derivation, which is a linear sequence of applying production rules. The co-occurrence of production rules is very helpful for avoiding rules that generate incorrect sentences [18]. The N-gram model of production rules predicts which subsequent production rule is likely to be used after particular rules are applied.

In this model, if the derivation is given by

$$S \xRightarrow{r_1} \gamma_1 \xRightarrow{r_2} \gamma_2 \xRightarrow{r_3} \cdots \xRightarrow{r_n} \gamma_n = x, \tag{3}$$

then the probability of this derivation is calculated as follows:

$$\begin{aligned} P(r_1, \ldots, r_n) &= P(r_1 \mid \#) P(r_2 \mid \#, r_1) \\ & \prod_{k=3}^{n} P(r_k \mid r_{k-2}, r_{k-1}) P(\# \mid r_n) \end{aligned} \tag{4}$$

In Equation 4, the rule sequence $r_1, \ldots, r_n$ is derived from top-down parsing. However, the LR parser is based on bottom-up parsing. Therefore, in this case, the probability of a derivation is defined as $P(r_n, \ldots, r_1)$.

In a context-free grammar, there is no constraint on the contexts in which a production rule can be applied. This model, however, is considered to have context-sensitivity in terms of probability. This is because the probability of a production rule is dependent on the previously applied rules.

Actually, the bigram model is used in continuous utterance recognition. We will give some consideration to the rule bigram model.

First, the rule bigram model can be considered as an extension of the word bigram model. Now consider the following CFG.

(R_1)	S	$\rightarrow$	$A\ B$
(R_2)	A	$\rightarrow$	$C\ D$
(R_3)	C	$\rightarrow$	w_1
(R_4)	D	$\rightarrow$	w_2
(R_5)	B	$\rightarrow$	w_3

In this case, we have the following rightmost derivation from the parsing of the string "$w_1w_2w_3$".

$$S \xRightarrow{R_1} AB \xRightarrow{R_5} Aw_3 \xRightarrow{R_2} CDw_3 \xRightarrow{R_4} Cw_2w_3 \xRightarrow{R_3} w_1w_2w_3 \tag{5}$$

From the viewpoint of bottom-up parsing, the rule bigram model must calculate the probability for the sequence R_3, R_4, R_2, R_5, R_1. Here, $P(R_4|R_3)$ is an approximation of the word bigram $P(w_2|w_1)$. Also, since $P(R_5|R_2)$ is approximated by $P(w_3|A)$, this probability is a word occurrence probability that takes the previous context into account. Thus, this model includes word bigram information.

Furthermore, the probability for the derivation is approximated by the following formula, which indicates that the rule bigram model is very similar to the *parse tree N-gram model* [19].

$$\begin{aligned} & P(R_3, R_4, R_2, R_5, R_1) \\ = \; & P(R_3|\#)P(R_4|R_3)P(R_2|R_4) \\ & P(R_5|R_2)P(R_1|R_5)P(\#|R_1) \\ \approx \; & P(w_1|\#)P(w_2|w_1)P(A|w_2) \\ & P(w_3|A)P(S|A)P(\#|S) \end{aligned} \tag{6}$$

5.3 Training of the Models

The stochastic language models were trained by actually parsing the sentences extracted from the *ATR dialogue database* [20].

Definition of Symbols

$\{B_1, B_2, \ldots, B_I\}$ $\cdots$ A set of training sentences.

$\{D^i_1, D^i_2, \ldots, D^i_{n_i}\}$ $\cdots$ A set of derivations for the i-th sentence B_i. Here, n_i represents the number of derivations for B_i.

$N^i_j(r)$ $\cdots$ The function N counts the number of rule occurrences of its arguments in the derivation D^i_j.

Training of the Stochastic CFG

The conditional probabilities of rules in the stochastic CFG were estimated using the following procedure [17].

1. Make an initial guess of $P(\alpha|A)$ such that $\sum_\alpha P(\alpha|A) = 1$ holds.
2. Parse the i-th sentence B_i and get the all derivations for B_i.
3. Re-estimate $P(\alpha|A)$ by the following formula.

$$\overline{P(\alpha|A)} = \frac{\sum_i C^i_A(\alpha)}{\sum_i \sum_\beta C^i_A(\beta)} \tag{7}$$

where

$$C^i_A(\alpha) = \sum_i \left(\frac{P(D^i_j)}{\sum_k P(D^i_k)} N^i_j(A \rightarrow \alpha) \right) \tag{8}$$

4. Replace $P(\alpha|A)$ with $\overline{P(\alpha|A)}$ and repeat from step 2.

Training of the Rule Bigram Model

The rule bigram probabilities were estimated by relative frequency of occurrence of production rules in derivations of training sentences. However, since ambiguous sentences have many different derivations, a simple relative frequency approch does not work well. The following formula was used to calculate the relative frequency, in which the rule frequency in one sentence is normalized by the number of derivations for that sentence.

$$P(r_n|r_{n-1}) = \frac{\sum_i \frac{1}{n_i} \sum_j N^i_j(r_{n-1}, r_n)}{\sum_i \frac{1}{n_i} \sum_j N^i_j(r_{n-1})} \tag{9}$$

6 PERFORMANCE EVALUATION

The HMM-LR speech recognition system was evaluated on a speaker-dependent conference registration task. This task consists of dialogues between secretaries and participants of international conferences.

Results of recognition experiments are shown in Table 3, where *Correct* indicates percent correct, and *Subs*, *Dels* and *Ins* are substitution error rate, deletion error rate and insertion error rate.

Compared with the stochastic CFG model, the rule bigram model attains very high recognition performance. Substitution and insertion error rates in particular decreased sharply. However, the deletion error rate is still high. This is because all recognition hypotheses are sometimes rejected in the middle of the input due to recognition failure at the very beginning. Indeed, the HMM-LR system gave no output for about 6~7% of all sentences. However, this phenomenon is not a shortcoming of the HMM-LR speech recognition system because, from the viewpoint of speech understanding, giving an incorrect result is much worse than no output.

Table 3: Recognition performance of the HMM-LR speech recognition system.

HMM	*Language*	*Word Recognition Performance*					*Sent. Recognition Performance*	
Training Data	*Models*	*Correct*	*Subs*	*Dels*	*Ins*	*Accuracy*	*Correct*	*Within top* 5
Word	CFG only	55.2	12.9	31.9	11.9	43.3	48.9	56.9
	Stochastic CFG	56.6	7.8	35.7	5.1	51.4	59.9	66.4
	Rule bigram	77.8	5.6	16.6	1.4	76.4	78.1	81.0
Word + Phrase	CFG only	66.7	12.5	20.8	6.8	59.9	55.5	64.2
	Stochastic CFG	69.1	9.2	21.8	3.3	65.7	64.2	72.3
	Rule bigram	83.1	2.9	14.0	0.1	83.0	83.2	86.9
Word + Phrase + Continuous	CFG only	70.8	13.5	15.7	6.9	63.9	57.7	65.0
	Stochastic CFG	74.1	10.5	15.4	4.1	70.0	66.4	71.5
	Rule bigram	81.0	2.9	16.1	0.0	81.0	83.9	86.1

7 CONCLUSION

This paper presented recent improvements in the HMM-LR speech recognition system. The system was evaluated through continuously spoken sentence recognition. These improvements include: (1) HMM training with continuous utterances as well as word utterances, (2) development of a sentential grammar, and (3) introduction of stochastic language models. The stochastic language models investigated were a stochastic CFG model and a bigram model of production rules. The recognition experiments demonstrated that the rule bigram model is much superior than the stochastic CFG. One reason for this is that the rule bigram model has context-sensitivity in terms of probability. The HMM-LR speech recognition system eventually attained a sentence recognition rate of 83.9% with the rule bigram model.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Tomita, M: "An Efficient Word Lattice Parsing Algorithm for Continuous Speech Recognition", Proc. ICASSP86, pp. 1569-1572 (April 1986).

[2] Saito, H. and Tomita, M.: "Parsing Noisy Sentences", Proc. COLING88, pp. 561-566 (August 1988).

[3] Kita, K., Kawabata, T. and Saito, H.: "HMM Continuous Speech Recognition Using Predictive LR Parsing", Proc. ICASSP89, pp. 703-706 (May 1989).

[4] Wright, J. H: "LR Parsing of Probabilistic Grammars with Input Uncertainty for Speech Recognition", Computer Speech and Language, Vol. 4, pp. 297-323 (1990).

[5] Kita, K. and Ward, W. H.: "Incorporating LR Parsing into Sphinx", Proc. ICASSP89, pp. 269-272 (May 1991).

[6] Goddeau, D. and Zue, V.: "Integrating Probabilistic LR Parsing into Speech Understanding Systems", Proc. ICASSP92, pp. I·181-I·184 (March 1992).

[7] Hanazawa, T., Kita, K., Nakamura, S., Kawabata, T. and Shikano, K.: "ATR HMM-LR Continuous Speech Recognition System", Proc. ICASSP90, pp. 53-56 (April 1990).

[8] Tomita, M. (Ed.): "Generalized LR Parsing", Kluwer Academic Publishers (1991).

[9] Sagisaka, Y., Takeda, K., Abe, M., Katagiri, S., Umeda, T. and Kuwabara, H.: "A Large-Scale Japanese Speech Database", Proc. ICSLP90, pp. 1089-1092 (November 1990).

[10] Kita. K., Takezawa T. and Morimoto T.: "Continuous Speech Recognition Using Two-Level LR Parsing", IEICE Transactions, Vol. E74, No. 7, pp. 1806-1810 (July 1991).

[11] Kita, K. and Morimoto, T.: "Language Models for Speech Recognition", JSAI Technical Report, SIG-SLUD-9201-9, pp. 59-60 (April 1992).

[12] Martin, W. A., Church, K. W. and Patil, R. S.: "Preliminary Analysis of a Breadth-First Parsing Algorithm: Theoretical and Experimental Results", In *Natural Language Parsing Systems*, Bolc, L. (Ed.), pp. 267-328, Springer-Verlag (1987).

[13] Su, K. Y., Chang, J. S.: "Semantic and Syntactic Aspects of Score Function", Proc. COLING88, pp. 642-644 (March 1992).

[14] Chitrao, M. V. and Grishman, R.: "Statistical Parsing of Messages", Proc. DARPA Speech and Natural Language Workshop, pp. 263-266 (June 1990).

[15] Magerman, D. M. and Marcus, M. P.: "Parsing the Voyager Domain Using Pearl", Proc. DARPA Speech and Natural Language Workshop, pp. 231-236 (February 1991).

[16] Black, E., Jelinek, F., Lafferty, J., Magerman, D. M., Mercer, R. and Roukos, S.: "Towards History-based Grammars: Using Richer Models for Probabilistic Parsing", Proc. DARPA Speech and Natural Language Workshop (February 1992).

[17] Fujisaki, T., Jelinek, F., Cocke, J., Black, E. and Nishino, T.: "A Probabilistic Parsing Method for Sentence Disambiguation", In *Current Issues in Parsing Technology*, Tomita, M. (Ed.), pp. 139-152, Kluwer Academic Publishers (1991).

[18] Kita, K., Kawabata, T. and Hanazawa, T.: "HMM Speech Recognition Using Stochastic Language Models", J. Acoust. Soc. Jpn. (E), Vol. 12, No. 3, pp. 99-105 (May 1991).

[19] Wrigley, A., Itou, K., Hayamizu, S. and Tanaka, K.: "Parse Tree N-grams for Spoken Lnaguage Modelling", JSAI Technical Report, SIG-SLUD-9202-12, pp. 104-113 (July 1992).

[20] Ehara, T., Ogura, K. and Morimoto, T.: "ATR Dialogue Database", Proc. ICSLP90, pp. 1093-1096 (November 1990).