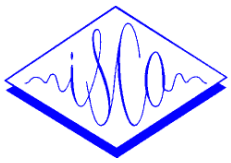

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

CONTINUOUS SPEECH RECOGNITION USING A DIALOG-CONDITIONED STOCHASTIC LANGUAGE MODEL

Hiroyuki Sakamoto and Shoichi Matsunaga

ATR Interpreting Telecommunications Research Labs.,
2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-02 Japan

ABSTRACT

This work attempts to improve recognition accuracy by predicting the next utterance in a dialog. We propose a dialog-conditioned stochastic language model that is applied to dialog speech recognition. Each dialog-conditioned stochastic language model has been constructed with text data of a certain situation and implemented using a Japanese syllable trigram. Each situation was defined in advance to predict next utterances effectively. Experiments in continuous speech recognition have shown that these models decrease perplexity and improve recognition accuracy in comparison with conventional dialog-uniform stochastic language models.

1. INTRODUCTION

Speech is a good way of achieving human-machine communications. It is generally used in dialog form. Therefore, it is suitable for language models that consider dialog forms in recognition systems. Uniform stochastic language models are usually used in speech recognition. Such models have achieved high performance in speech recognition experiments involving bigrams or trigrams (e.g.[1][2][3]). In the future, however, we expect various situations to be represented in dialog form and language information to be directed toward individual situations. In human-machine dialog systems, we hope that it will be effective to use language models according to individual situations. In many cases, it is possible to predict a subject's utterances. For example, if the machine asks a "yes/no question", the subject's answer usually begins with "yes/no". Therefore, if the machine uses a dialog-conditioned stochastic language model instead of the conventional uniform model, it should be able to achieve a more accurate performance.

It was previously reported that in a limitative task domain, the words of a user's utterance are restricted as the dialog progresses and can therefore be predicted. It was also stated that by considering the sentential type for the system's utterance, the syntax of the user's utterance can be predicted [4].

This paper proposes a dialog-conditioned stochastic language model and uses Japanese syllable trigrams to compare its performance with that of a uniform model.

2. DIALOG-CONDITIONED STOCHASTIC LANGUAGE MODEL

2.1. Clustering Text Data

Dialog-conditioned stochastic language models are generated by text data for each situation in dialog form. Therefore, a dialog-conditioned stochastic language model can be considered an appropriate language model for a particular situation. Representative situations are determined on the assumption that the next utterance can be predicted in the dialog and that text data can be categorized for individual situations. If we assume the domain "registration at an international conference", a participant's utterances (about 4.7×10^4 phrases) can be classified into several situational categories.

These situational categories are given as the following:

- START (*opening of dialog*)
- WHO (*asking subject's name*)
- WHAT (*asking subject's business*)
- YES/NO (*asking yes or no*)
- OTHER (*a situation other than the above situations*)

In all of these situations except for OTHER, it is possible for the participant's utterances to be predicted by the utterances of a member of the conference's secretariat.

2.2. Generating Dialog-Conditioned Stochastic Language Model

The dialog-conditioned stochastic language models are then generated from the classified text data. This study employed a Japanese syllable trigram as the stochastic model, and the training text data for the Japanese syllable trigram included 108 symbols (107 syllables and a symbol for silence). Japanese syllables consist of either a vowel or a combination of a consonant and a vowel. To compensate for any lack of text data, syllable trigram models are normalized by deleted interpolation[5]:

$$\tilde{P}(w_n|w_{n-2},w_{n-1}) = \lambda_0 P_0 + \lambda_1 P(w_n) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n|w_{n-2},w_{n-1}) \quad (1)$$

where $\tilde{P}(w_n|w_{n-2}, w_{n-1})$ is the normalized trigram model, and w represents a syllable. The term $P(w_n|w_{n-2}, w_{n-1})$ is the probability of syllable w_n occurring after the syllable sequence w_{n-2}, w_{n-1}, and λ is a combination factor such that $\sum_{i=0}^{3} \lambda_i$ is equal to 1.

We propose that a trigram model P^* should have more probability of accuracy than a trigram model $\tilde{P}$ generated from the classified text data. The trigram model P^* is described by

$$P^* = \eta\tilde{P} + (1-\eta)\tilde{P}_{rest} \quad (2)$$

where $\tilde{P}_{rest}$ is generated from the text data excluding the classified text data to generate $\tilde{P}$, and η and (1-η) are the mixture weights of the interpolated model represented by $\tilde{P}$ and $\tilde{P}_{rest}$.

Figure 1 shows a dialog system that uses the proposed dialog-conditioned stochastic language model. In this system, "dialog control" generates system utterance and predicts the next user utterance. Next, "speech recognition" recognizes user utterance with the selected dialog-conditioned stochastic language model based on the prediction, and its results are reported to "dialog control".

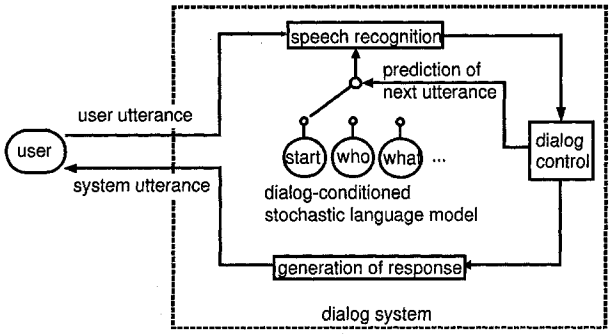

Figure 1: Dialog system using dialog-conditioned stochastic language model

2.3. Perplexity with Dialog-Conditioned Stochastic Language Models

2.3.1. *Perplexity on Conference Registration Task*

Table 1 shows the syllable perplexity of the dialog-conditioned stochastic language model. (··) is the number of phrases.

For 321 test-set phrases concerning conference registration, the average syllable perplexity (6.27) of the dialog-conditioned stochastic models is less than that (8.43) of the random stochastic language models generated from randomly-selected data of the same text, which contain the same amount of data as the dialog-conditioned models. It is also less than the perplexity (6.90) of a uniform stochastic language model generated from all data. The lowest average syllable perplexity (6.02) was achieved by the models generated by equation (2).

2.3.2. *Task Dependency of Dialog-Conditioned Model*

Another task we investigated was "travel arrangement" (about 7×10^3 phrases). Table 2 shows the syllable perplexity for the START situation, WHO situation, WHEN situation (*asking subject's time*) and for all test-set phrases (for comparison). (··) is the number of phrases.

Table 1: Perplexity of $\tilde{P}$ and P^*(conference registration)

training text		test phrase	perplexity
START	(4110)	START	5.3
random	(4110)	(74)	5.9
START*			4.7
WHO	(460)	WHO	9.1
random	(460)	(61)	33
WHO*			10
WHAT	(4377)	WHAT	7.9
random	(4377)	(26)	8.2
WHAT*			7.0
YES/NO	(5121)	YES/NO	7.6
random	(5121)	(54)	8.2
YES/NO*			6.9
OTHER	(32515)	OTHER	5.04
random	(32515)	(106)	5.36
OTHER*			5.14
uniform	(46583)	all (321)	6.90
average	condition		6.27
	random		8.43
average*			6.02

For 1019 test-set phrases concerning travel arrangement, the average syllable perplexity (12.2) of the dialog-conditioned stochastic models is less than that (15.5) of the random stochastic language models generated from randomly-selected data, which contain the same amount of data as the dialog-conditioned models. This result shows that the dialog-conditioned models do not depend on the task.

2.3.3. *Supplementing Dialog-Conditioned Stochastic Language Model*

In Table 2, the syllable perplexity (10.8) of a uniform stochastic language model generated from all data is less than that of the dialog-conditioned stochastic models. This is because the amount of training text data is insufficient. However, the models generated by equation (2) produce the lowest perplexity (10.3) among these models. Thus, the models generated by equation (2) are suitable, when the amount of training text data is insufficient. For the conference registration task, a similar tendency can be observed.

Table 2: Perplexity of $\tilde{P}$ and P^*(travel arrangement)

training text		test phrase	perplexity
START	(306)	START	9.6
random	(306)	(67)	18
START*			6.9
WHO	(424)	WHO	14
random	(424)	(53)	40
WHO*			14
WHEN	(333)	WHEN	12
random	(333)	(32)	40
WHEN*			11
uniform	(7091)	all (1019)	10.8
average	condition		12.2
	random		15.5
average*			10.3

2.3.4. *Discussion on Using the Utterance just before Predicted Utterance*

We expect the predicted utterance to be affected by the utterance just before it. For example, if an utterance is "What is your name?", the next utterance can be predicted in all likelihood to begin with "My name is ...". Therefore, we have investigated using the previous utterance for prediction of the next utterance. Text data for classified participant's utterances is added to text data for a secretariat member's utterances to enable the prediction of the next participant's utterance. The language models generated by this text data include the utterance just before the next utterance.

Table 3 shows the perplexity of these models for the START, WHO and WHEN situations in a travel arrangement task. It can be seen that similar results are observed for similar situations. [··] is the number of additional times corresponding to the case with the lowest perplexity.

Table 3: Perplexity of models; perplexity with added utterance just before the predicted utterance

test	training text	
START	START	START+[2]
	9.55	9.54
WHO	WHO	WHO+[1]
	13.7	13.6
WHEN	WHEN	WHEN+[4]
	12.1	11.3

These additional models produce little effect compared with the dialog-conditioned models. The reason for this is that the utterance just before the next utterance has very little text data. However, the WHEN situation, in particular, has a significant effect. In the WHEN situation, the utterance just before the next utterance explains an appointed day, and the next utterance is largely affected by the contents of the utterance just before the predicted utterance. We confirm the possibility of making good use of immediately previous utterances in some types of situations.

3. SPEECH RECOGNITION EXPERIMENTS

In this paper, the effect of the dialog-conditioned model was examined from the viewpoints of recognition rate using a task-oriented recognition system and recognition rate using a grammar-free dictation system. The dialog-conditioned model was also evaluated using a speech recognition system both when the task-oriented grammar was and was not used in a conference registration task. Furthermore, speaker-dependent speech recognition experiments were conducted using Japanese phrases uttered by one male.

The experimental conditions are summarized in Table 4.

The speech recognition experiments employed a hidden Markov network (HMnet) [6], which is an efficient representation of a phoneme context-dependent HMM. A single Gaussian 600-state HMnet was trained with 2620 Japanese words uttered by one male speaker. Both the Baum-Welch algorithm and Vector Field Smoothing (VFS) [7] were used for the training. Since the VFS algorithm has a smoothing procedure, we could decrease the amount of training data.

Table 4: Experimental Conditions

Analysis conditions	
Sampling rate	12 kHz
Window	Hamming window (20 ms)
Frame period	5 ms
Analysis	log power + 16-order LPC-Cep + Δlog power + 16-order ΔLPC-Cep
Training data	
Speaker	1 male (MHT)
Samples	2620 Japanese words
Adaptation and recognition data	
Speaker	1 male (MTK)
Adaptation data	25 words
Recognition data	321 phrases

We experimented on an SSS-LR continuous speech recognition system. This SSS-LR system was combined the HMnet, derived by using a Successive State Splitting (SSS) algorithm, with a phoneme-context-dependent LR parser [8].

The method of calculating the score by this system is given as

$$P_s = (1-\theta)\log P_h + \theta \log P_l \qquad (3)$$

where P_s is the total score, P_h is the score of the acoustic model by the HMM models, P_l is the score of the stochastic language model (e.g. dialog-conditioned stochastic language model, uniform model), and θ is the P_h and P_l weight.

3.1. Task-Oriented Grammar Recognition

In phrase recognition experiments, the dialog-conditioned stochastic language models were combined with a generalized LR parser [8], which can cope with a context-free grammar. The grammar included 1321 vocabulary words and 2813 rules.

Table 5 shows the recognition results for each of the situations (e.g. START, WHAT, WHO) and for all situations (all) when the conference registration grammar was used.

#OTHER is the result of the recognition experiments using the uniform model generated by all of the text data; in this case, this is useful because the participant's utterances could not be predicted by the secretariat member's utterances. "Average" denotes the average of the recognition results obtained for each situation when the uniform model was used in the OTHER situation. (··) is the number of phrases.

Phrase recognition rates were as follows: 82.2% when the dialog-conditioned model was used, 80.7% when the random model was used, and 81.0% when the uniform model was used. If we assume insufficient training text data, it is probable that models generated by equation (2) would provide greater accuracy. However, the amount of training text data in these experiments was sufficient,

Table 5: Phrase recognition rate (%) using a task-oriented grammar (θ=0.5)

rank	test phrase	stochastic language model used		
		condition	random	condition*
top 1	START	74	70	73
top 5	(74)	85	82	84
top 1	WHAT	73	73	73
top 5	(26)	81	77	77
top 1	WHO	85	79	80
top 5	(61)	89	87	85
top 1	YES/NO	87	89	89
top 5	(54)	91	91	91
top 1	#OTHER	85.9		
top 5	(106)	92.5		
top 1	average	82.2	80.4	81.3
top 5	(321)	88.8	87.5	87.5

rank	test phrase	uniform model
top 1	all	81.0
top 5	(321)	87.2

as indicated by the phrase recognition rate obtained using models generated by equation (2) being 81.3% (less than the 82.2% achieved by the dialog-conditioned model). Without the stochastic language model ($\theta = 0$), the phrase recognition rate was 76.0% (top5: 86.0%). Therefore, the dialog-conditioned stochastic language model produced the best phrase recognition rates.

3.2. Grammar-Free Dictation Recognition

Next, in grammar-free dictation recognition, we studied the effect of using only the stochastic language model by applying phonotactic constraints apparent in Japanese. Table 6 shows the results of phrase recognition experiments obtained for each situation when the grammar was not used.

Table 6: Phrase recognition rate (%) without using a grammar (θ=0.8)

rank	test phrase	stochastic language model used		
		condition	random	condition*
top 1	START	55	51	58
top 5	(74)	70	62	74
top 1	WHAT	38	31	38
top 5	(26)	50	42	54
top 1	WHO	52	18	49
top 5	(61)	57	23	57
top 1	YES/NO	63	57	50
top 5	(54)	74	67	78
top 1	#OTHER	67.0		
top 5	(106)	79.3		
top 1	average	58.6	49.5	56.4
top 5	(321)	69.8	59.5	71.7

rank	test phrase	uniform model
top 1	all	55.5
top 5	(321)	69.8

Phrase recognition rates were as follows: 58.6% when the dialog-conditioned model was used, 49.5% when the random model was used, and 55.5% when the uniform model was used. Thus, we found that the dialog-conditioned stochastic language model again produced the best phrase recognition rates in comparison with other language models. We therefore expect improvement in the recognition rate by using the dialog-conditioned stochastic language model.

4. CONCLUSIONS

This paper has proposed a dialog-conditioned stochastic language model and has examined the effect of the model on speech recognition. A secretariat member's utterances were used as a standard against which we could cluster the participant's utterances concerning registration at an international conference. Accordingly, we could generate a dialog-conditioned stochastic language model. We performed experiments on phrase recognition by using this model in individual situations. For both task-oriented grammar recognition and grammar-free dictation recognition, the recognition rate was improved by using the dialog-conditioned model. Based on this, we concluded that the proposed method achieved a more accurate speech recognition performance than other methods. However, this method may produce the opposite effect when the participant's utterances are different in predicted situations because this method is completely changed in only predicted situations.

In the future, automatic classification methods for these situations will be studied. In addition, we plan to study the dialog-conditioned stochastic language model as a Japanese word n-gram model.

REFERENCES

[1] Kai-Fu Lee, "Large-Vocabulary Speaker Independent Continuous Speech Recognition: The SPHINX System," 15213 CMU-CS-88-148 (April 18, 1988).

[2] A. Averbuch, et al., "An IBM PC Based Large-Vocabulary Isolated-Utterance Speech Recognizer," Proc. of ICASSP'86, Vol.1, 2.4.1 pp.53-56 (1986).

[3] T. Araki, J. Murakami and S. Ikehara: "Post-processing in Japanese Speech Recognition Using 2nd-order Markov Model of Syllables," NTT REVIEW, vol.1, no.3, pp.96-104 (1989.9).

[4] Y. Moriya, T. Abeno, M. Yamamoto and S. Nakagawa: "A Spoken Dialog System with Prediction of the Next User Utterance," technical report of IEICE, SP92-121, pp.43-50 (1993.1). (In Japanese)

[5] F.Jelinek: "The development of an experimental discrete dictation recognizer," Proc. IEEE, vol. 73, pp.1616-1624 (1985).

[6] J. Takami and S. Sagayama: "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. of ICASSP'92, pp. 573-576 (1992).

[7] K. Ohkura, M. Sugiyama and S. Sagayama: "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. of ICSLP'92, We.fPM.1.1, pp. 369-372 (1992).

[8] A. Nagai, S. Sagayama, and K. Kita: "Phoneme-context-dependent LR Parsing Algorithms for HMM-based Continuous Speech Recognition," Proc. of Eurospeech'91, S48.3. (1991).