MODELING OF SPOKEN DIALOGUE WITH AND WITHOUT VISUAL INFORMATION

Katsuhiko SHIRAI

Department of Information and Computer Science, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169 Japan
E-mail: ks@shirai.info.waseda.ac.jp

ABSTRACT

Auditory information is the major factor in the human communication but in practical conversations, visual information such as gesture, facial expression, and head movement clearly makes it much smoother and more natural. Most researches related to analysis of spoken dialogue are based on only auditory information. We are trying to clarify how human utilize the knowledge of spoken dialogue management by dealing with more natural communication that includes visual information. Above all, by using visual information we can deal with listener's attitude against the speaker that cannot be done by using only auditory information.

1. INTRODUCTION

Human beings communicate with one another by conversation using speech and speech is regarded as the most natural means of communication for humans. This can be said for communication between humans and computer. The progress of the fundamental technologies including speech recognition and spoken language processing in recent years, gives the high performance of speech understanding. However, the systems that can talk with humans naturally is not constructed up to now. To construct such a system, it is necessary to analyze how humans behave and what phenomenona happen when communicate with one another by speech.

To realize a natural dialogue like those between humans on the computer, it is necessary to make use of not only acoustic information but also the linguistic knowledge and more high-level knowledges of dialogue management. The collection and analysis of natural dialogue corpora is the main method of clarifying how humans actually use such knowledge.

From this point of view, spoken dialogue corpora for various tasks have been collected and analyzed and models are made and applied for spoken dialogue systems.

In this paper, recent topics on collection of Japanese dialogue corpora and the efforts to analyze them are introduced in Section 2 and in Section 3, and a dialogue model based on corpora is discussed in Section 4. In Section 5, the effects of visual information channel is discussed.

In practical conversations, visual information such as gesture, facial expression, and head movement are important factors to make it much smoother. Most researches related to analysis of spoken dialogue are focused on only auditory information channel. We made a preliminary experiment to clarify how much the dialogue management differs if it includes visual information. Through the visual channel, both attitudes of the listener and the speaker that may not be shown by using only auditory channel necessarily can be represented.

2. COLLECTION OF DIALOGUE CORPORA

Statistical methods based on large corpora are the main method applied to recent speech recognition and linguistic processing.

Therefore, the importance of speech database is understood widely and many research laboratories have been collecting Japanese spoken dialogue corpora. However, Japanese spoken dialogue corpora are still in want and so much more speech database is desirable.

In various scenes of their daily life, humans are communicating by speech, but several considerations have occurred. To collect spoken dialogue, it is necessary to set up the situation of conversation and to make experiment on collecting simulated spoken dialogue. The contents of conversation are influenced by the situation. They will be changed by who is talking on what kind of tasks in what situation. Therefore, the selection of tasks and setting up of situation are the important problem.

For example, Kurematsu has been collecting and analyzing the characteristics of spoken dialogue corpora on the scheduling task[1] and Tsuchiya has been making a database of spoken dialogue corpora on Japanese map task[2].

3. ANALYSES OF SPOKEN DIALOGUE CORPORA

Not only linguistic information but also acoustic characters such as intonation fill the important roles in humans' communication by speech. From this point of view, Ichikawa has

proposed the method of description of pitch patterns on dialogue and the method of estimation of construction of an utterance in semireal time[3]. Their objective is the construction of a dialogue understanding system which is capable of real-time processing by estimating all sorts of roles of conversational phenomenona such as turn taking and stumbling from dialogue corpora with pitch patterns. And Ueda is trying to propose a guide to a plan of spoken dialogue corpora by estimating characteristics and contents of dialogue by analyzing acoustic information such as phonemes and moras quantitatively[4].

4. DIALOGUE MODELS WITHOUT VISUAL INFORMATION

To construct a system that can generate natural and flexible response for spoken dialogue system, a dialogue model including various of phenomenona appeared in humans' conversation.

Doshita and Araki proposed a dialogue model that reflects two important aspects of spoken dialogue system: to be 'robust' and to be 'cooperative'[5]. For the purpose, their model has two inference functions: conversational function and problem solving function. The conversational function is a kind of dynamic Bayesian network that represents a meaning of utterance and general dialogue rule. The problem solving function is a network so called Event Hierarchy that represents the structure of task domain problems. To construct conversational function and estimate in problem solving function, they divide the dialogue processing from speech understanding through generating response into 5 stages:

1. understanding of meaning
2. intention understanding
3. renewal will condition
4. intent generation
5. generation of response

Kita has proposed an ergodic HMM and a dialogue model by means of a state merging method for the corpora with a speaker label and an utterance type called IFT[6].

And our viewpoint is that many analyses have been done on dialogues, but such analyses are made independently on dialogues based on one single task but not on several tasks. Therefore, we propose a non-task-oriented dialogue model based on the dialogue corpora for several tasks[7]. We conceive a non-task-oriented dialogue model by analyzing corpora based on several tasks and extracting the utterances which are not task-oriented.

4.1. Dialogue corpora

We analyzed the following four dialogues in Simulated Spoken Dialogue Corpus(SSDC)[8].

- Task 1: crossword puzzle task[7].
  Two subjects have across or down hints of same crossword puzzle, and exchange each knowledge in order to accomplish the puzzle.
- Task 2: scheduling task[1].
  Two subjects set the date for a meeting with each schedule tables.
- Task 3: telephone shopping task[4].
  Two subjects play a customer or a clerk. The clerk asks the customer some information for the shopping.
- Task 4: reservation guide task[9].
  Two subjects play a travel agent or a customer. The customer applies a tour.

Table 1 Ratio of task-oriented utterance.

	task 1	task 2	task 3	task4	all
Non-task-oriented	836	132	122	195	1285
Task-oriented	694	53	61	110	918
Total	1530	185	183	305	2203
Task-oriented/Total	45%	29%	33%	36%	42%

Functional labels which show functional roles in the dialogue were attached to all utterances in the corpus. It might be happened that one utterance had several labels.

We could classify these utterances to two types. One was task-oriented utterances which conveyed indispensable information of the task. For example, in crossword puzzle task hints or answers were task-oriented utterances. The other was non-task-oriented utterances which did not relate to the task. For example, greeting or acknowledgment were non-task-oriented utterances.

Table 1 shows ratio of task-oriented utterance. We found out that from 60% to 70% of all utterances were not task-oriented, by analyzing them in terms of their roles in the dialogues based on several tasks. Furthermore, the non-task-oriented utterances were classified into labels related to turn-taking.

Table 2 shows list of label of non-task-oriented utterance.

Table 2 Labels of non-task-oriented utterance

Label		Turn-Taking
TS	Topic-Start	unsettled
TP	Topic-Pause	unsettled
TE	Topic-End	unsettled
IA	Information-Addition	utterer
TR	Information-Renewal	utterer
IL	Information-Lack	utterer
RS	Recognition-Success	companion
RR	Recognition-Repeat	companion
RF	Recognition-Failure	companion
CA	Contents-Affirmation	companion
CC	Contents-Confirmation	companion
CD	Contents-Denial	companion

4.2. Dialogue Model

By making a Markov model of those labels that are common to several tasks, we made a non-task-oriented model. Figure 1 shows 5 state ergodic HMM in the condition that both state transition probabilities and output probabilities are over *0.1*. Then we examine the propriety of the labels by comparing the model with the Markov model of characteristic labels which are obtained from the words in utterances.

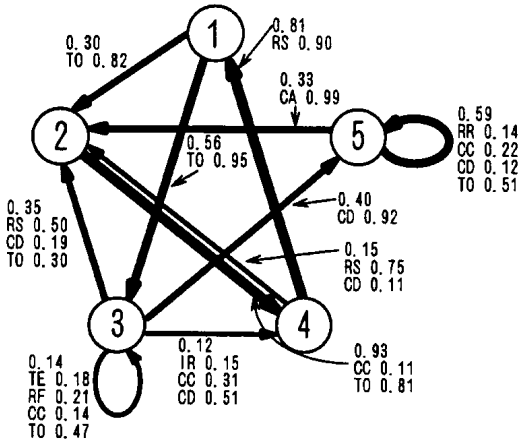

Figure 1 5 state ergodic HMM for labels of non-task-oriented utterance.

4.3. Discussion

From the model in Figure 1, we can see that the state ① corresponds to a start of a topic, the state ② corresponds to an end of recognition leading the end of a topic, the state ③ corresponds to exchanging information of the task, and the state ④ corresponds to an affirmative end and ⑤ corresponds to a confirmation of recognition or content.

5. DIALOGUE MODELS WITH VISUAL INFORMATION

Certainly, auditory information is the major factor in the human communication, but in practical conversations, visual information such as gesture, facial expression, and head movement clearly makes it much smoother and more natural. Here we make a new dialogue modeling with visual information. As such a information, humans head movements such as face-up, vertical head movement when he/she say "No", horizontal head movement in saying "Yes" and so on are used. In Subsection 5.1, head movement labels are described and then a new model with visual information was built on HMM in Subsection 5.2.

Table 3 Head movement label

Label		Speaker's frequency	Listener's frequency
F	Face up and look his partner	179	45
V	Vertical movement of head	54	30
H	Horizontal movement of head	5	1
I	Inclined movement of head	11	4
O	Other movements	96	92
N	No movement	603	776

5.1. Head Movement Label

We chose task 1 in Subsection 4.1 for analysis but in this case, two subjects are faced and can see each other. As the nature of crossword most of the time subjects are facing down looking at the clues, it was hard to take a video suitable for automatic extraction of head movements. Therefore we decided to extract them by hand[11].

To make the flow of the dialogue clear, we labeled the utterance according to auditory and visual information. For the label of visual information, we defined 6 labels. As the nature of crossword puzzle most of the time subjects are facing down, therefore we defined facing down as a normal action and prepared a label for when the subject faced up to watch his partner. And for a certain utterance subjects can be divided into two situations that are speaker and listener, so we labeled separately. In Table 3, we show the labels and the number of appearance about the speaker and the listener. We allowed more than one utterance in a sentence to label. The total number of the labels were 948.

5.2. Visual Information Model

The chosen dialogues have 3 different labels, speech function, speaker's head movement and listener's head movement. Figure 2 shows 5 state ergodic HMM for labels of non-task-oriented utterance in the case that two subjects are faced and can see each other, and Figure 3 for 3 different labels.

In both models, both state transition probabilities and output probabilities are shown if they are over *0.1*.

5.3. Discussion

Comparing Figure 1 and Figure 2, we can see the difference between dialogues in two different conditions that two subjects are faced and can see each other or not. Because those two models are different, we can say that by using visual information we can deal with listener's attitude against the speaker that cannot be done by using only auditory information.

From Figure 2 and Figure 3, we can see that the model with visual labels has more state transitions to itself than the model without them so the model with visual labels has a

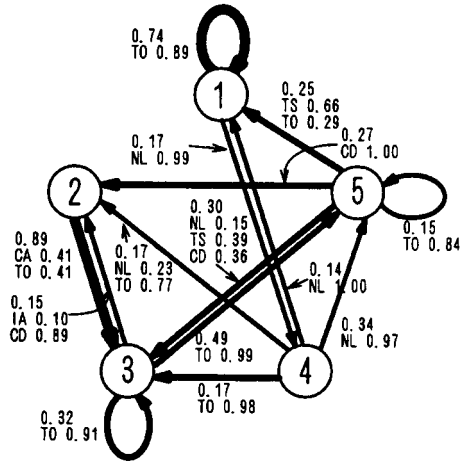

Figure 2 5 state ergodic HMM with visual information but without head movement label.

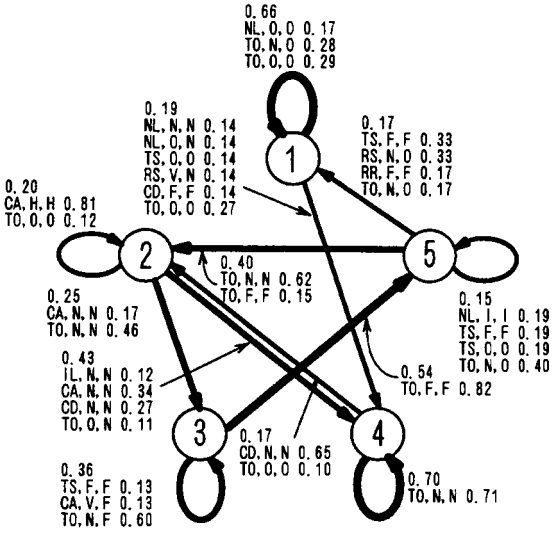

Figure 3 5 state ergodic HMM with head movement labels.

tendency to be settled.

6. CONCLUSION

To apply to man-machine interface using speech, many analyses and modeling have been done on dialogues, and many dialogue systems based on dialogue models have been made. However, no system realize a natural dialogue such as the dialogues between human applied those models up to now. Moreover, in some case, information can be obtained from the movement of head easily that cannot be done from only auditory information. Therefore visual information is effective to have more natural conversation.

In the coming multimodal dialogue system, the use of visual information will become important and both auditory and visual information might have to be modeled together. From now on, it will become important problem how to apply these models to man-machine interface.

7. REFERENCES

[1] Kurematsu A. and Nakasuji T.: "Characteristics of Spontaneous Speech on the Scheduling Task", *1996 Spring National Convention IEICE*, SD-4-2, pp.329-330, 1996, (*in Japanese*).

[2] Tutiya S. and Koiso H.: "Utterance and Word Units and Their Representaions in the Chiba University Map Task Dialog Corpus", *Technical Report of JSAI*, SIG-SLUD-9501-4, pp.25-32, 1995, (*in Japanese*).

[3] Ichikawa A., et al: "Description of Estimation on Dialogue ", *Proc. Spring Meet. Acoust. Soc. Jpn.*, 1-P-4, pp.169-170, 1996 (*in Japanese*).

[4] Ueda N. and Itahashi S.: "Characteristics of High Occurrence Frequency N-gram of Spoken Japanese Corpora", *Proc. Spring Meet. Acoust. Soc. Jpn.*, 1-P-20, pp.201-202, 1996 (*in Japanese*).

[5] Araki M. and Doshita S.: "Cooperative Spoken Dialogue Model using Bayesian Network and Event Hierachy", *Trans. of IEICE*, Vol. E78-d, No. 6, pp629-635, 1995.

[6] Kita K., et al: "Automatically Deducing Probabilistic Dialogue Models from an IFT-Annotated Corpus", *Technical Report of JSAI*, SIG-SLUD-9503-8, pp47-54, 1996 (*in Japanese*).

[7] Morikawa E., et al: "Analysis of Utterance Based on Corpora for Several Tasks", *1996 Spring National Convention IEICE*, SD-4-1, pp.327-328, 1996 (*in Japanese*).

[8] Grant-in-Aid for Scientific Research on Priority Areas Project: "Research on Understanding and Generating Dialogue by Integrated Processing Of Speech, Language and Concept– Simulated Spoken Dialogue Corpus Vol.1", 1994.

[9] Ikeda Y., et al: "Topic prediction based on utterance motivation", *Proc. Autumn Meet. Acoust. Soc. Jpn.*, 2-10-3, pp.89-90, 1995 (*in Japanese*).

[10] Nishimoto T., et al: "Improving Human Interface in Drawing Tool Using Speech, Mouse and Key-board", *Proc. of ROMAN95*, pp.107-112, Jul. 1995.

[11] Iwano Y., et al: "Analysis of head movements and its role in spoken dialogue" *Proc. of ICSLP96*, 1996.