

ANALYSIS OF HEAD MOVEMENTS AND ITS ROLE IN SPOKEN DIALOGUE

Yuri Iwano, Shioya Kageyama, Emi Morikawa, Shu Nakazato and Katsuhiko Shirai

School of Science and Engineering,
Waseda University

ABSTRACT

In this research, we analyzed the relationship between semantics of utterances and movements of head in a natural dialogue and a task oriented one in Japanese. We are going to show that visual information such as head movement will be useful for managing a dialogue and reducing the vagueness of semantics. First we extracted the head movements calculated automatically in a natural conversation and going to indicate the role of it. After this we will show an analysis of head movements during a cooperative problem solving task to construct a natural dialogue system in which initiative of the conversation moves. We will show the effectiveness of using visual information in a multimodal dialogue system.

1. INTRODUCTION

Dialogue is an interactive communication of information mainly based on speech. For instance, considering conversation using a telephone, we can have natural conversations without actually seeing each other. However, in practical conversations, visual information such as gesture, facial expression, and facial movement clearly makes it much smoother and more natural. Unlike speech, non linguistic behavior have no standard rule and differences among individuals are large. Among these human movements in a conversation, the head movement is clear and it is easy to count, so it is one of the visual information that we can grasp it objectively. Using the information not only from a speech signal but also knowledge of linguistic characteristics of spoken dialogue and dialogue management method is important for the accomplishment of natural dialogue on a computer. Most researches related to analysis of spoken dialogue are based on only auditory information. We are trying to clarify how human uses 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 a listener's attitude against the speaker that cannot be done by using only auditory information.

The research done here is all based on Japanese language. It is said that during conversation, Japanese speakers give more responses like nodding than English speakers. And the way we answer questions are different from English. We always

decide to give affirmative or negative answer according to the question not the fact. Therefore, the result of this research is culturally bounded although we still can tell the importance of visual information.

2. ANALYSIS OF NATURAL DIALOGUE

2.1. Experiment

To analyze the head movements in a conversation, we did an experiment under a natural circumstance. We recorded one of the subject's conversations on a VCR for approximately one hour. We chose 444 parts of the conversation that had clear head movements, and labeled according to the function of the utterance and the movement of head.

2.2. Extraction of head movement

We placed a video camera between two subjects who were facing each other and recorded their conversation without obstructing their view. Using a workstation that has a video input capability, we captured the conversation into it. The movement of head is calculated automatically in vertical, horizontal and inclined directions. Vertical and horizontal directions were calculated by searching the position of the nose and inclined direction by the relation of the position of two eyes(Figure 1).



Figure 1: Extraction of head movements

2.3. Labels of the dialogue

We defined labels for the visual information according to a combination of vertical, horizontal and inclined directions and its amount of movements. In Table 1 we show these

labels and the number of appearances. The movement large and small are classified by a threshold and movement none includes very small movements that could be an error range. The vertical movements appeared the most followed by inclined and horizontal direction.

Table 1: Number of appearances of head movement

label	inc.	ver.	hor.	frequency
A Inclined L	L	*	*	27
B Vertical L	#	L	*	170
C Horizontal L	#	#	L	35
D Inclined S	S	N	N	23
E Vertical S	N	S	N	103
F Horizontal S	N	N	S	7
G Inc.Ver. S	S	S	N	16
H Inc.Hor. S	S	N	S	2
I Ver.Hor. S	N	S	S	1
J All S	S	S	S	1
K No movement	N	N	N	59

L-large, S-Small, N-None, *-Any, #-Small or None

We now label the dialogues in a view point of function of utterances that we already have the head movement’s labels. The speech function’s labels are based on who has the speech turn and the function of utterance. Table 2 shows the labels of speech function and their number of appearances.

Turn-taking “companion” means, leaving the speech-turn to his companion or giving away speech turn to his companion. Turn-taking “speaker” means, the speaker obtained the speech turn and told some sort of information.

Table 2: Speech function’s labels and its frequency

turn-taking	labels	frequency
1 Unsettled	Topic-start	1
2 Speaker	Topic-pause	0
3 Unsettled	Topic-end	0
4 Speaker	Information-addition	1
5 Speaker	Information-renewal	3
6 Speaker	Information-lack	4
7 Companion	Recognition-success	206
8 Companion	Recognition-repeat	5
9 Companion	Recognition-failure	31
10 Companion	Contents-affirmation	94
11 Companion	Contents-denial	80
12 Companion	Contents-confirmation	19

2.4. Relation between speech functions and head movements

From Table 2 as we chose 444 places of dialogue according to whether the head moved or not we can tell the followings.

- When turn-taking is “companion”, utterances involve many head movements.

- When turn-taking is “speaker” or “unsettled”, utterances involve few head movements.
- When turn-taking is “companion” and the function of speech is “recognition-success”, utterances involve many head movements.
- Utterances related to contents involves many head movements.

In Table 3 we show the relation between labels of speech functions and head movements. Columns represent the head movement’s labels in Table 1 and rows represent the speech function’s labels in Table 2.

Table 3: Relation between speech functions and head movements

head move	Speech function’s label number												total
	1	2	3	4	5	6	7	8	9	10	11	12	
A	0	0	0	0	1	2	10	0	2	3	8	1	27
B	0	0	0	1	0	0	102	0	4	51	4	8	170
C	0	0	0	0	0	0	0	0	0	0	35	0	35
D	0	0	0	0	0	1	5	4	5	0	6	2	23
E	1	0	0	0	0	0	61	0	4	28	5	4	103
F	0	0	0	0	1	0	0	0	0	0	6	0	7
G	0	0	0	0	0	0	7	0	3	0	4	2	16
H	0	0	0	0	0	0	0	0	0	0	2	0	2
I	0	0	0	0	0	0	0	0	0	0	0	1	1
J	0	0	0	0	0	0	0	0	0	0	1	0	1
K	0	0	0	0	1	1	21	1	13	12	9	1	59
total	1	0	0	1	3	4	206	5	31	94	80	19	444

By paying attention to the distinct relation between labels of speech functions and head movements in Table 3 that is when turn-taking is companion and speech functions are “recognition-success(7)”, “contents-affirmation(10)” and “contents-denial(11)”, we can say the followings.

- Vertical head movements(B, E) represent “recognition-success(7)” and “contents-affirmation(10)”, and it also represents its degree.
- When the speech function is “contents-affirmation(7)” and it involves head movements, its direction is vertical(B, E) or inclined(A, D).
- The horizontal movements(C, F) clearly represent “contents-denial(11)”, but it does not seem to show its degree.
- If the speech function is not “contents-denial(11)”, having a horizontal head movements(C, F) are rare.

When head inclines(A, D), it occurred often when the speech functions are “recognition-success(7)” and “contents-denial(11)”. However, we did not have a category, we could say the following points.

- When head inclines, the speech function seems to be “recognition-success” with withholding feelings and “contents-denial” with skeptical feelings.

Among the total of 444 data, the frequency was very low but we confirmed some exceptional data that had the opposite result from what we have said in the previous part of this paper.

- When the speech function is contents-denial in a negative content of dialogue, head could move vertically.

3. DIALOGUE IN COOPERATIVE PROBLEM SOLVING

We have discussed about the movements of head in a natural dialogue. Here we are going to analyze the head movements in a dialogue that has a definite task. Analysis of dialogues with a task would give more information for constructing a dialogue system between man and machine than one that has not. Therefore we are going to choose crossword puzzle solving as a task that is a cooperative problem solving task dealt previously by the group of Nakazato and Morikawa [3][4][5]. In dialogue analysis we can expect more detailed analysis by adding visual information. We analyzed the dialogue including the movements of the listener that cannot be dealt with only by auditory information.

3.1. Experiment

We asked two subjects to cooperate to solve a crossword puzzle. They are given a clue either vertical or horizontal direction. Two subjects cannot see each other's clue but they are faced and can see each other. We recorded two subjects' movements and their voices by using two video cameras. We did not assign any limitation on their conversations. They can speak whatever they want. The experiment will finish when both subjects' crosswords are filled completely.

We did an experiment with four different crosswords and the total number of sentences was 778. As the nature of a crossword most of the time subjects are facing down looking at the clues, taking a video suitable for automatic extraction of head movement was hard. Therefore, we decided to extract visually.

3.2. Labels of the dialogue

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 categories. As the nature of a 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. For a certain utterance, subjects can be divided into two situations, speaker and listener. We labeled these two situations separately. In Table 4 we show the labels and the number of appearances of the speaker's head movement and in Table 5 the listener's. We allowed labeling more than one utterance in a sentence. The total number of the labels was 948.

Table 4: Head movement's labels of speaker

label	frequency
Face up and watch his partner	179
Vertical movement of head	54
Horizontal movement of head	5
Inclined movement of head	11
Other movements	96
No movement	603

Table 5: Head movement's labels of listener

label	frequency
Face up and watch his partner	45
Vertical movement of head	30
Horizontal movement of head	1
Inclined movement of head	4
Other movements	92
No movements	776

For the function of speech we used 24 labels, and it is shown in Table 6.

Table 6: Number of appearances of speech functions

labels	frequency
Hail to partner	70
Giving response	27
Indicate topic (location)	195
Indicate question (clue)	192
Answer	116
Agree with answer(with knowledge)	22
Objection	7
Ask question	42
Ask again	9
Confirm answer	17
Confirm location	31
Confirm clue	46
Agree with answer	1
Affirmation to the question	39
Correction of answer	13
Talk to oneself (thinking)	9
Talk to oneself	27
Objection to the previous unit	3
Insufficient information	19
Skeptical agreement	8
Proposal of solving strategy	14
Agreement to the strategy	2
Repeat	5
The others	34

3.3. The relation between dialogue structure and head movement

It is known that in a crossword puzzle task, the whole dialogue from the beginning to the end, consists of several units that are small section to derive 1 vertical or horizontal crossword puzzle answer[5].

We analyzed the relation between length of the unit and the head movement. The 4 dialogues in this experiment contained total of 146 units and the average number of labels during one unit was 6.49. Furthermore in Table 7 we calculated the average probability of having a head movement for 3 different lengths of the unit (The number of labels having head movements in the unit divided by the total labels in that unit).

Table 7: Relation between length of unit and head movement

Number of labels per unit	Probability of involving head movements
20 or more labels	40.4%
10 to 19 labels	34.6%
1 to 9 labels	26.3%

As the result we can say when the length of the unit gets longer, in other words, when it takes longer time to lead one answer, it shows a tendency to have more head movements.

It seems when exchanging the information to solve the crossword puzzle is going well, the auditory information is enough but once it is not, we request more information through visual information, such as facing his head up to see his partner's reaction.

3.4. Relation between speech functions and head movements

The chosen 948 dialogues have 3 different labels, speech function, speaker's head movement and listener's head movement. Here we analyzed what kind of combination rise frequently. We omitted the dialogue with no head movement labels and listed the high appearance probability (probability to have the written head movements among the same speech functions) combinations in Table 8.

Table 8: Relation between speech functions and head movements

contents	prob. (%)	frequency
Affirmation	43.59	17
Vertical mov. (speaker)		
Giving response	29.63	8
Vertical mov. (speaker)		
Agree with answer	27.27	6
Vertical mov. (speaker)		
Ask question	26.19	11
Face up (speaker)		
Speak to partner	24.29	17
Face up (speaker)		
Indicate question	22.92	44
Face up (speaker)		
Confirm clue	17.39	8
Face up (speaker)		
Indicate topic	10.77	21
Face up (speaker)		
Question	9.48	11
Vertical mov. (listener)		
Indicate question	4.17	8
Face up (listener)		

From Table 8 we can say the followings.

- Affirmation, agreement and giving responses involve vertical movement of head.
- When the speaker wants to have a response from the listener, speaker often faces up to see his partner.
- When listener moves his head vertically, he is giving an affirmative response to the speaker.

4. CONCLUSION

We analyzed the relation between the speech function and head movement in a dialogue of a natural and cooperative problem solving task. We showed the role of the head movements in a spoken dialogue.

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.

Considering a man-machine interface, when the conversation is going smoothly, interface with auditory information can handle the situation but once it becomes complicated, visual information as a feedback to the system to show the user's situation will be effective. In the coming multimodal dialogue system, the use of visual information will become very important.

The information we can retrieve from visual information is not only head movement. For instance, there are expressions, gestures, a glance and so on. We are going to include these information to our analysis. Also we want to analyze the difference between the dialogue that had a visual contact with his partner and not.

5. REFERENCES

- [1] Senko K. Maynard: "Kaiwa Bunseki", Kuroshio publisher, 1993 (in Japanese)
- [2] Keiko Watanuki, Fumio Togawa: "Some signals of emotional arousal: Analysis of conversations using a multimodal interaction database", Eurospeech 95, pp.1165-1168, 1995.
- [3] Shu Nakazato, Shuichi Tanaka, Katsuhiko Shirai: "Roles of utterances in cooperative problem solving", Proc. spring meeting Acoust. Soc. Jpn., 3-P-18, pp.197-198, 1995. (in Japanese)
- [4] Shu Nakazato, Emi Morikawa, Katsuhiko Shirai: "Analysis of utterances functions in cooperative problem solving", Proc. autumn meeting Acoust. Soc. Jpn., 1-Q-29, pp.197-198, 1995. (in Japanese)
- [5] Emi Morikawa, Shu Nakazato, Shuichi Tanaka, Katsuhiko Shirai: "A dialogue model for cooperative problem solving task", Proc. of the 9th Annual Conference of JSAI, 19-01, pp.525-528, 1995. (in Japanese)