
4-3 Interaction Analysis at the Dialog by Nonverbal Behavior

YOSHIMOTO Jun, MIZUKAMI Etsuo, YAMASHITA Koji, and YANO Hiroyuki

We much know nonverbal behaviors that coexist with our communicative history may tell us important information and regulate our verbal dialogue when we have face to face conversed. Researches for nonverbal behaviors in human interaction were begun by scientific ways, however, quite recently. In order to measure the activity of conversation by the computer with video cameras, it is necessary to detect regulators that called turn-maintaining cues or turn-yielding cues. This paper presents a method of automatic segmentation and classification of nonverbal behaviors in dialogues captured on video from two subjects by cluster analysis. As an example for evaluating a classification, we discuss the specific group of behaviors had suggested turn-maintenance cue.

Keywords

Nonverbal behavior, Regulator, Turn-maintaining

1 Introduction

Aiming to address the challenges inherent in the diversity and ambiguity of nonverbal behavior, we will first describe the role of such behavior in face-to-face communication in the introduction below, followed by a discussion of our findings based on related research.

1.1 Three layers of face-to-face communication

What sort of structure is formed by the information we transmit and receive when we attempt to establish communication with another person? It is generally assumed that in face-to-face communication, this structure consists of three layers.

First, there is the language layer, which consists of words arranged in an order that conforms to grammatical rules. This layer forms the core of the information to be communicated, and its formation allows information to be relayed to and shared with another

person. This is the layer employed when we perform general activities such as reading and writing, and the facility with which the information within this layer may be converted into text form is one of the main features that characterizes this layer.

Next, there is the paralinguage layer. This layer accompanies the speech act performed by the language layer. In other words, this layer corresponds to that portion of voice information created by speech minus the language layer, and includes the pitch and volume of the voice as well as the pacing and pauses of speech. Using voice communication tools such as telephones, multi-layered information consisting of the language and paralinguage layers may be exchanged reciprocally. Generally, since the speech act requires no tools and is easier to learn relative to the act of writing, the physical load placed on the user is smaller. The paralinguage layer enables modification of the language layer and permits the speaker to relay his intentions and describe his state; this layer is therefore essential to high-

modality communication. However, information contained in this layer outside the language layer contains a certain degree of ambiguity, and so information consciously transmitted by the speaker may not be received, or the audience may attach undue importance to an involuntary movement of the speaker. (For example, the speaker may have intended to emphasize an item of information by speaking loudly, but the audience may have heard the item as spoken softly; or the speaker may believe himself to be speaking at normal speed, while the audience is under the impression that he is speaking rapidly, erroneously concluding that he is in a hurry.) Compared to words, which have common meaning within a single linguistic or cultural area, paralinguage by definition entails uncertainty in meaning or intention and potential misperceptions caused by intra-speech and inter-speech relativity, as in the examples given above. Paralinguage is also inconsistent between individuals and difficult to convert into text. Thus, although recognized as an important layer of communication, the paralinguage component has not been actively pursued as a theme of past studies. Further, since the ability to use the paralinguage layer is acquired by learning based on actual communications, significant differences are found among speakers from different regions. For example, major differences are clear among dialects in Japan, not only in vocabulary but also in intonation, including accent. These differences, among others, render it difficult to implement voice recognition applications for various provincial groups.

Finally, there is the nonverbal layer, which consists of visual information relayed during face-to-face communication. This nonverbal layer may be said to convey information even through the appearance and clothing of the speaker; however, in the present paper, we will mostly deal with the body actions of a person while speaking or listening. One theme of nonverbal studies is the concept of pose, which addresses the pairing of patterns of expression in the paralinguage layer (so-called "elocution") and the nonverbal action

layer. However, as stated above, this paper deals mainly with actions. Compared to the paralinguage layer, the nonverbal layer is complex and covers a wide range of information. Thus, it is an extremely difficult layer to decode and even more difficult to convert into text format. Nevertheless, in some cases the nonverbal layer can nearly match the language layer in significance. Sign language, mainly used by the hearing or speech challenged, serves as an alternative to the speech act as a means of expression. It is also possible to reverse the meaning of a vocal speech act by a single facial expression. For example, the phrase "Thank you" may be spoken calmly, yet a cold expression, clenched fists, and tense shoulders reveal an intent that is at odds with the literal meaning of the phrase. Furthermore, the presence of regulators that support fluent turn-taking in communication proves that this layer is as important as the paralinguage layer in communication and interaction. However, despite sporadic past studies, this area has not formed the focus of continuous research.

1.2 Background of nonverbal studies and issues to be resolved

It must be said that we lack a detailed understanding of nonverbal behavior. Such behavior has probably formed a part of human speech throughout history, but it was not until the late 19th century that the first study on gesture and greetings was initiated, and formal studies on the effects of nonverbal behavior on communication had to wait until the late 20th century. For example, the first book on nonverbal communication in the US appears to have been published in 1972. (Details on the history of nonverbal studies may be found in [1] and [2].) A videophone that allowed communications incorporating nonverbal information was displayed at the 1970 Japan World Exposition, but to this day, the general public has yet to fully enjoy video communications that would accommodate nonverbal content, despite the passage of decades since. The development of technology to convey nonverbal information in communications has been the subject of

continuous research, but it remains unclear in what manner such nonverbal information is employed in different situations, or what effects it has on the audience. Thus we must undertake to assess the conditions surrounding the use of nonverbal elements and attempt to elucidate the cognitive mechanisms involved, as well as to clarify the cumulative effect of nonverbal elements on the communication in question. Efforts must also be devoted to the development of technologies related to the above-mentioned research as regards application to various systems.

1.3 History of research on the classification of nonverbal behavior

The concept of body language[3] is now recognized widely not only among researchers but also among the general public. However, the rules concerning nonverbal language deviate from those of spoken language. Based on a structural approach within the field of kinesics, Birdwhistell[4] has attempted detailed identification and classification of nonverbal behavior using classification methods similar to those used in linguistics. He proposed new concepts such as the allokine, kine, and kineme, which correspond to the concepts of allophone, phone, and phoneme in linguistics. His methodology assumed that, just as multiple and meaningless microsounds are connected to form significant speech, microbehaviors having no meaning are connected to form an action with significance. However, a number of counterarguments have been proposed[5][6], and Birdwhistell's methodology has not become widely adopted among researchers. In short, attempts to classify nonverbal behavior using linguistic classification methods remain problematic.

In contrast, the argument has been made that nonverbal behavior cannot be classified without taking into consideration the specific context in which such behavior arises. Ekman and Friesen have proposed an approach using external variables for kinesics that does not attempt a structural classification of nonverbal

behavior, but instead classifies behaviors into several basic categories based on the purpose, meaning, and intent of each[7]-[11]. This approach is now widely embraced among researchers in the field.

1.4 Present paper's approach to classification of nonverbal behavior

The strong points in the classification schemes of Birdwhistell and Ekman and Friesen introduced in the previous section have been applied to design a classification scheme for nonverbal behavior. First, video recordings of communication sessions are made and behaviors exceeding a certain threshold of activity are identified. Next, the identified behaviors are compared with one another in a round-robin fashion based on the expressed paralanguage layer. Similar behaviors are grouped into a given category. Up to this point, this method resembles that of Birdwhistell, and thus presents the same shortcomings. However, for behaviors featuring simple structures, it is assumed that these shortcomings may be ignored. Some major problems in the prior structural approach were clear: the behaviors were broken up too finely, and there are no clear breaks in these behaviors, such as those found in speech. Behaviors cannot be easily segmented. Ultimately, categories having characteristics corresponding to the nonverbal behavior classification scheme proposed by Ekman and Friesen were carefully selected in order to assess the validity of the present scheme.

We believe that it is worthwhile to attempt a classification of nonverbal behavior as part of basic studies on nonverbal communications. A communication experiment will be described in Section 2 below, with a presentation of a method of analysis for the obtained data in Section 3.

2 Communication experiment

2.1 Preparation

In order to acquire basic data on the role of nonverbal behavior in communication, we conducted an experiment in which two subjects were asked to form a pair, videotaping communications between the subjects as they performed given tasks. Each subject was placed in a room surrounded by walls that prevented each from seeing or hearing the other. The interaction between the subjects was carried out using a video camera, monitor, microphone, and headphones. A video camera was set up in front of each subject, so both subjects were able to have a frontal view of each other as they communicated. Video cameras to record side views were also installed. The video images and voice data recorded by these instruments were first stored and then later input to a computer for analysis. The recorded video data consisted of the frontal and side views of both subjects. Additionally, since the subjects were placed in separate rooms, their speech acts were recorded independently on different channels.

2.2 Analysis strategy

The video image was recorded in NTSC format at a rate of 29.97 fps. Behavioral expressions on extremely short time scales are likely to be overlooked by the other party, and so were considered to have inherently low significance. In addition, the latency between the time a certain action is seen or heard and the moment a reaction is expressed is normally 200–400 ms. Taking these two points into consideration, the frame rate selected for the analysis was 7.49 fps (133 ms per frame), which was believed to be the most appropriate period in which to observe the propagation and simultaneous occurrence of behaviors in the subject interactions.

3 Data processing

3.1 Automatic segmentation of nonverbal behavior

For each frame of the video, a region to be measured was set for each subject. The total sum of the difference in brightness from frame to frame for the given region was measured in order to calculate the subject's amount of movement per frame. When the amount of movement failed to exceed a certain threshold, the subject was considered to have remained still. Figure 1 shows an example of an image in a frame recorded in the experiment. The grids in the side view show the target region for behavior detection. Thus, the subject is judged to have moved when the brightness within this region was subject to change. A single behavior was defined as a period of movement continued over a preset length of time, separated from other behaviors by periods in which no movement was observed in respect to the time axis — i.e., the still condition.

Top and bottom panels show the frontal and side views of both subjects, respectively. The grids on the bottom panel indicate the target region for movement detection.

Figure 2 shows an example of the nonverbal behavior chart created by automatic segmentation of behaviors expressed during a 241-second communication session between a given pair of subjects. The amount of movement is represented logarithmically on the vertical axis, and the horizontal axis represents



Fig. 1 Example of recorded video image

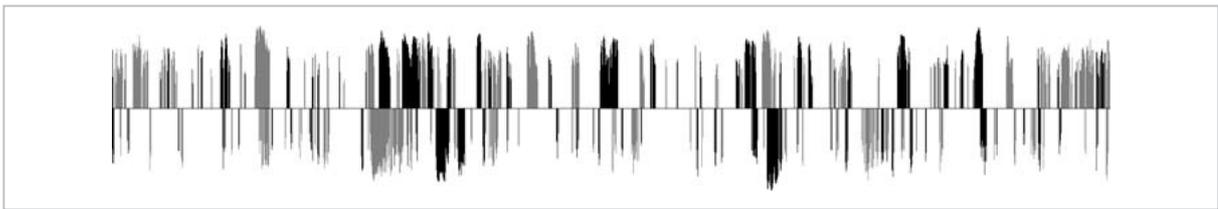


Fig.2 Nonverbal behavior chart

time. Each behavior is given an identification number, with alternating shadings for even- and odd-numbered behaviors for each subject, to render visual presentation more readily comprehensible.

Top and bottom panels show the behaviors of subjects A and B, respectively, with the vertical and horizontal axes representing the amount of movement and time, respectively.

While only behavior is shown in Fig. 2, the volume of the voice in the corresponding speech act is also shown on the same time axis in Fig. 3, and a system was developed to construct a video image with synchronized voice data using the frontal video recording of the subjects. This approach enabled us to obtain an overview of the status of communications both in the still and video images.

The top plot signifies the volume of speech, similarly to the plot in Fig. 2. The bottom panel shows the part of the body where movement has occurred based on light intensity, for the frontal view of the subject.

3.2 Classification of nonverbal behavior by cluster analysis

According to the proposal by Ekman et al., the segmented behavior must be carefully examined individually in order to classify the behavior into one of five categories: “emblem”, “illustrator”, “regulator”, “affect display”, and “adopter” [7]-[12]. However, these procedures were not followed in the present study. Instead, we have attempted a schematic and automatic classification based on the characteristics of the respective behaviors. In the initial step of our analysis, the duration, amount of movement, and volume of speech during the behavior for each behavior segment were used as variables (for an initial

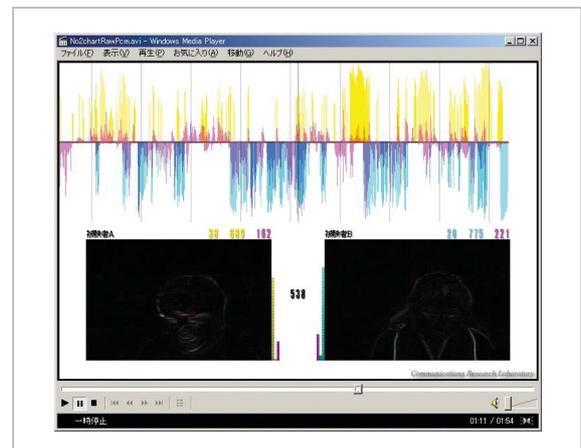


Fig.3 Overview of the communication session

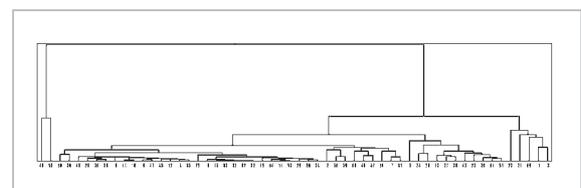


Fig.4 Dendrogram for classifying head movement

variable group). Next, each variable was standardized to unify the units involved, followed by cluster analysis of the relevant behavior for each subject. Figure 4 shows a dendrogram obtained as a result of a typical UPGMA cluster method [13] performed on the 56 behaviors detected in a given subject. When the individual behaviors and behavior clusters are connected at the lower positions, then the behaviors may be regarded as highly similar. Conversely, when connected at the higher positions the behaviors are more dissimilar. In the left half of Fig. 4, many behaviors connected at the lower positions are observed, indicating that these are similar behaviors.

The behavior is labeled by an identification number. Behaviors connected at lower

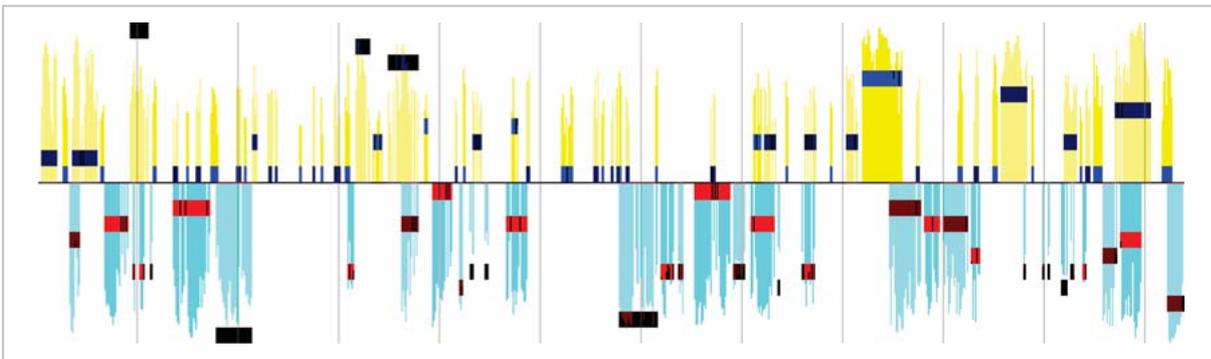


Fig.5 Example of nonverbal chart with classification codes

positions are highly similar, while those connected at higher positions are considered more dissimilar.

Figure 5 is an example of a nonverbal chart with classification codes for communication between a subject pair for a 114-second communication session. As in Fig. 2, the amount of movement is shown in log scale in the vertical axis for each subject, and time is represented by the horizontal axis. The classification codes showing the groups to which the respective behaviors belong is also shown with rectangles on the plot. The number of groups may be changed by altering the height at which the dendrogram obtained by cluster analysis is divided. In this case, behaviors were categorized into 10 qgroups. Those displaying high similarities in initial variable groups — in other words, similar behaviors — were placed in the same group. For a single subject, the behaviors marked by rectangular symbols located at the same height belong to the same group. This is analogous to the notation used in musical scores. Just as notes at the same height in the score show notes at the same pitch, rectangular symbols at the same height reflect behaviors in the same group. This analogy is common; in addition to the presentation here^[14], a number of other studies have expressed results using plots similar to musical scores^[15].

Using the plot in Fig. 2, classification codes are presented to show the group to which the behavior belongs using the height of the rectangular code. In the top plot for subject A, the behavior with the code nearest to the

central axis represents behavior observed mainly as nodding.

3.3 Examination of the classified behavior

Behavior reflecting nearly the same pose at the beginning and end of the behavior is defined as a “close-ended” behavior, and one having different poses at the endpoints is defined as an “open-ended” behavior. An open-ended behavior is created by the intervention of a pause in the behavior, and actually belongs to a series of long-duration behaviors although this series initially appears to be discontinuous. Since each such open-ended behavior is in fact divided into two behavior units under the present method, only close-ended behaviors will be discussed below. Methods of analysis are available that can handle open-ended behaviors, but these methods were not applied here.

When performing cluster analysis, the precision of classification depends on the variable group provided. However, this does not mean that higher precision can be achieved by using larger numbers of variable groups. In contrast, better results may be obtained by intentionally narrowing the number of variables based on the type of behavior for which higher precision is required. The initial variable group selected for classifying behaviors schematically is suitable for apparent behavior classification, for example, but is not suitable for the classification of regulators.

Thus, a variable group for regulators was selected and consisted of the duration of

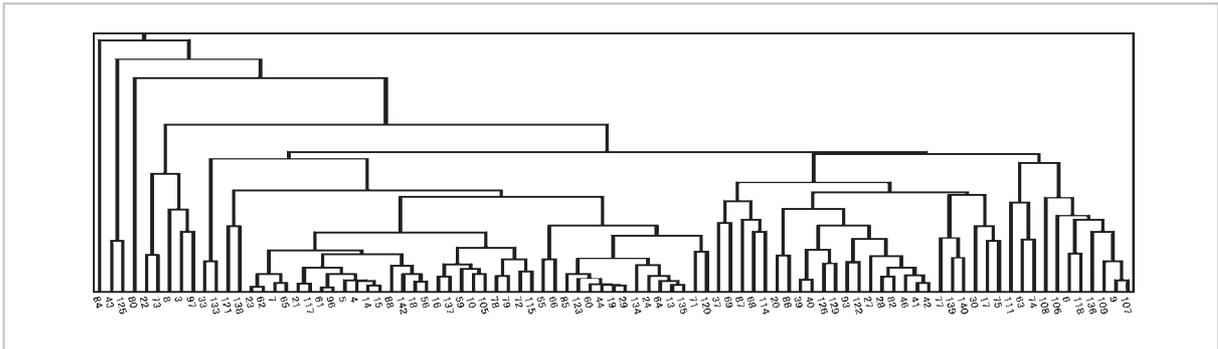


Fig.6a Dendrogram showing the classification of head movement only for the close-ended behavior

behavior, the average relative speech ratio during the behavior of each subject, the average independent speech ratio during the behavior of each subject, and the volume of speech during the behavior. Figure 6a shows the dendrogram obtained by cluster analysis using this variable group performed on a subject pair, and an enlarged section of the tree for one group is shown in Fig. 6b. The validity of the regulator classification will be examined using these results.

The format is the same as in Fig. 4, except that the result of cluster analysis is limited to 81 close-ended behaviors out of a total of 145 behaviors (whole view).

This example shows the results for a 251-second communication session between two Japanese subjects. A total of 145 behaviors were observed, out of which 81 were close-ended behaviors. Of these, eight behaviors (# 6, 9, 106, 107, 108, 109, 118, and 136) were selected and compiled in Table 1 together with the contents (language layer) of the communication immediately preceding them. (As stated in Section 2.1, the subject pair was given a take, with video and voice recording of their interaction as they solved the problem together. The speech content shown in Table 1 follows the flow of their efforts. In this case, the two subjects are looking at a photo of two people, and are trying to determine which person is the manager and which is the shop clerk [16].) During the behavior, the subject under observation performs speech acts such as “Hmm” or “Er”, and judging from the con-

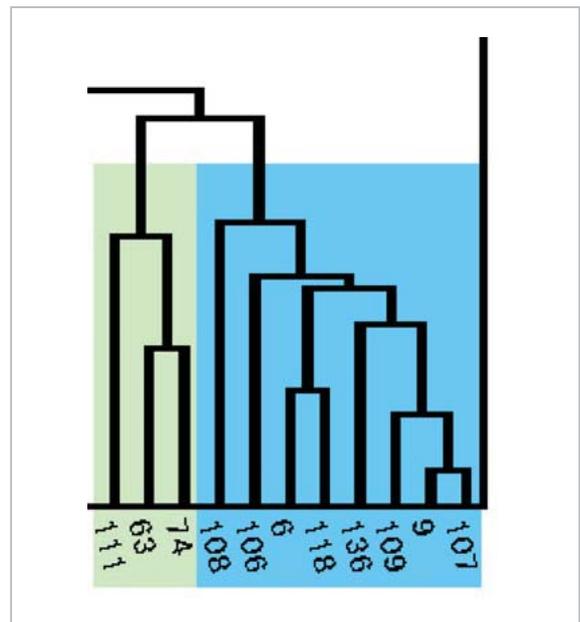


Fig.6b Dendrogram (enlarged view)

tent of the speech performed by the partner immediately before the behavior, it is clear that the observed behavior categorized into the group shown in Fig. 6b is the regulator for the turn-maintenance.

4 Conclusions

We have shown that it is possible to perform automatic segmentation and classification of behaviors by selecting appropriate variables for highly similar behaviors featuring simple structures in the classification scheme proposed by Ekman et al., while incorporating the structural aspect proposed by Birdwhistell. Using this method, we were

Table 1 *Content of speech associated with the behavior of the observed subject*

Behavior#	Duration (frames)	Partner's speech content immediately preceding behavior	Speech act of the observed subject
6	20	There is no evidence	Hmm
9	24	Yes	Er
106	12	I get the feeling person #1 is	Hmm
107	24	Manager of the shop	Hmm
108	23	I think he's talking about something	Hmm
109	32	And person #2	Hmm
118	22	The manager is in the back, far back	Hmm
136	37	Now, person #1 is	Hmm

able to identify the status of communication based on characteristics of voice and behavior through observation of communication between a pair of subjects engaged in basic human interaction, using a computer for subsequent analysis.

To improve the precision of the identification of the communication and to obtain a deeper understanding of the states of the two subjects during their communications, analysis of the language and paralanguage layers will be required. Although not addressed in the present paper, a number of studies were also

performed in parallel with this study: an examination of agreement expressions in cooperative tasks[17], a study on modeling of the appropriateness of speech manners between two people based on the rhythm of pauses and voice power[18], and a study on the identification and classification of fillers and emotional exclamations as mental markers for analyzing the mental state of the speaker[19][20]. The results of these studies have also contributed to increasing our understanding of communication.

References

- 1 A. Kendon, "Gesture: Visible Action as Utterance", New York; Cambridge University Press, 2004.
- 2 V. P. Richmond and J. C. McCroskey, "Nonverbal behavior in Interpersonal Relations", Allyn and Bacon, 2003.
- 3 R. L. Birdwhistell, "Kinesics and Context: Essays on Body Motion Communication", Philadelphia: University of Philadelphia Press, 1970.
- 4 R. L. Birdwhistell, "Introduction to Kinesics: An Annotation System for Analysis of Body Motion and Gesture", Louisville, KY: University of Louisville Press, 1952.
- 5 A. T. Dittmann, "Review of kinesics and context by R. L. Birdwhistell", *Psychiatry*, 34, 34-342, 1971.
- 6 V. P. Richmond, "Nonverbal Communication in the classroom", Acton, MA: Tapestry Press, 1996.
- 7 P. Ekman, "Movements with precise meanings", *Journal of Communication*, 26, 14-26, 1976.
- 8 P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception", *Psychiatry*, 21, 88-106, 1969.
- 9 P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding", *Semiotica*, 1, 49-98, 1969.

-
- 10 P. Ekman and W. V. Friesen, "Hand movements", *Journal of Communication*, 22, 353-374, 1972.
 - 11 P. Ekman and W. V. Friesen, "Detecting deception from the body or face", *Journal of Personality and social Psychology*, 29, 288-298, 1974.
 - 12 K. R. Scherer and P. Ekman, "Handbook of Methods in Nonverbal Behavior Research", New York; Cambridge University Press, 1982.
 - 13 H. C. Romesburg, "Cluster Analysis for Researchers", Florida: Robert E. Krieger Publishing Company Inc., 1989.
 - 14 J. Yoshimoto and H. Yano, "Classification of Head Movements at Video Captured Dialogue by Cluster Analysis", *Proceedings of the 2004 General Conference of Institute of Electronics, Information and Communication Engineers*, Mar. 2004. (in Japanese)
 - 15 M. Nishiyama, H. Kawashima, T. Hirayama, and T. Matsuyama, "Facial Expression Representation based on Timing Structures in Faces", *IEEE International Workshop on Analysis and Modeling of Faces and Gestures* (W. Zhao et al. (Eds.): AMFG 2005, LNCS 3723), pp.140-154, Oct. 2005.
 - 16 D. Archer, "How To Expand Your S.I.Q.(Social Intelligence Quotient)", New York: M. Evans and Company Inc., 1980. (Japanese translation: T. Kudo, E. Ichimura, "Body-language Kaidokuhoh", Tokyo: Seishinshobo Co. Ltd., 1988.)
 - 17 H. Yano and J. Yoshimoto, "Analysis of Verbal / Nonverbal Information of Agreement Expression in Mutual Agreement Dialogue", *Special Interest Group on Spoken Language Understanding and Dialogue Processing*, SIG-SLUD-A203-07, 41-46, 2003. (in Japanese)
 - 18 E. Mizukami, "How the Conversational Rhythm of 'MA' can be Constructed in Japanese Dialogue", In *Proceedings of The 8th World Multi-Conference on Systemics, Cybernetics and Informatics*, 14, 3-8, 2004.
 - 19 E. Mizukami, K. Yamashita, and H. Yano, "Effects of Modality and Familiarity on Dialogue to Describe a Figure: Analysis of Speech Fillers", *Progress in Asian Social Psychology Series*, 6, 343-358, 2007.
 - 20 K. Yamashita and E. Mizukami "Shinteki-marker ni yoru shinteki-syori process no rikai — Zukei setsumei kadai taiwa ni okeru filler wo cyushin to shita bunseki —", *Journal of Natural Language Processing*, 14(3), 39-60, 2007. (in Japanese)



YOSHIMOTO Jun, Ph.D.

Researcher, Spoken Language Communication Group, Knowledge Creating Communication Research Center (Former: Researcher, Social Interaction Group, Keihanna Human Info-Communication Research Center, Information and Network Systems Department)

Computer Science, Nonverbal Communication



MIZUKAMI Etsuo, Ph.D.

Former: Researcher, Social Interaction Group, Keihanna Human Info-Communication Research Center, Information and Network System Department

Filler, Multi-Party Interaction, Interaction Analysis



YAMASHITA Koji, Ph.D.

Former: Researcher, Social Interaction Group, Keihanna Human Info-Communication Research Center, Information and Network Systems Department

Cognitive Psychology, Educational Technology, Communication (Nonverbal, Media)



YANO Hiroyuki, Dr. Eng.

Planning Manager, Strategic Planning Section, Strategic Planning Division (Former: Group Leader, Social Interaction Group, Keihanna Human Info-Communication Research Center, Information and Network Systems Department)

Dialogue Process, Cognitive Model of Dialogue