My name is Kazuhiro Otsuka from NTT Communication Science laboratories. Today I would like to talk about our research activity. The title is "Communication Scene Analysis Based on Probabilistic Modeling of Human Gaze Behavior".
This shows the outline of my talk.

First, let me begin with a short introduction of our laboratory, and describe the motivation behind our research, which is the problem with current visual communication system.

Next, as a solution to the problem, I will show you "t-Room", that is an immersive teleconferencing system.

Then, I will move on to the conversation scene analysis, today's main topic, including inferring conversation structure based on gaze behavior and quantifying interpersonal influence in conversation.

In addition, I will briefly introduce the issue of visual representation, which is an automatic video editing using gaze pattern.

Finally, let me conclude my talk.
The NTT Communication Science Laboratories is one of 12 laboratories of NTT, Nippon Telegraph and Telephone Corporation, and its research areas includes human science and computer science such as learning theory, information theory, media processing, language processing, and so on.

The goal of our laboratory is overcoming communication barriers.
One motivation behind our research comes from the location of our laboratory, which is split among two sites; one is in Atsugi, and the other is in Kyoto, they separated by about 300 miles.
Motivation

Current video conferencing system has problems...

Atsugi Kyoto

Communication barriers exist

...Difficult to know who is talking to whom

Here, a communication barrier exists in terms of distance. We often use a video conferencing system as shown here to have meetings. But, we encounter problems with current video conferencing systems, as shown here.

For example, people at Kyoto look at this display, and they feel it difficult to know who is talking to whom. Actually, this person is talking to him, but it is not easy to understand that.
Motivation

Current video conferencing system has problems…

Instead of dividing screen, how about this? In this picture, maybe you can see they are facing each other and having a conversation. However, people’s faces are too small to distinguish who is talking and even who is who in this screen.

…Difficult to know who is talking to whom
Looking for Solutions

- Hardware Solution
  - Multiple large LCD panels
  - Multiple cameras and microphones
  - Broadband network ~ 100Mbps
  - Create immersive environment shared with people at different locations

- Software Solution
  - Flexible and portable devices
  - Less expensive
  - Narrowband network ~ 1Mbps
  - Create social reality by intelligent computer mediation.

To solve the problems with current communication system, we have been looking for solutions from two aspects; one is so called hardware side and the other is from software side.

The hardware solution employs multiple LCD panels, multiple cameras and microphones, and broadband network to create immersive communication environment that can be shared with people at different locations.

On the other hand, the software solution is something that requires less expensive devices such as PC, PDA, and cell phone with relatively narrow band links. But, we aim to create social reality by building computer-mediated communication system.
As a hardware solution, we have been developing a system called “t-Room”. Indeed, I am not in charge of this, but on behalf of Dr. Hirata and people developing the system, let me give you an outline. At first, let me show you a video clip to explain how this works.

[Video clip]

As you can see in this video, participants can feel as if people at different places are in the same place, because the geometric inconsistency of users’ gaze directions is partially resolved by a special configuration of displays and cameras.

You can feel as if people at different places are in the same room.

Geometric inconsistency of gaze directions is (partially) resolved by a special configuration of displays and cameras.
Here, let's quickly review this configuration.

The image of a person at place A is captured by this camera, and transmitted to this display at place B. In a same manner, a person at place B is taken by this camera and displayed on this screen at place A. Here, the important thing is that the viewing angle of cameras is set to correspond to display area, so that person's figure is displayed in the same position, at the same size, at the other place. Also, each camera uses a polarized filter so as not to capture the images displayed on the LCD, which would otherwise create a feedback loop.
Looking for Solutions

Hardware Solution
- Multiple large LCD panels
- Multiple cameras and microphone
- Broadband network ~ 100Mbps
- Create immersive environment shared with people at different locations

Software Solution
- Flexible and portable devices
- Less expensive
- Narrowband network ~ 1Mbps
- Create social reality by intelligent computer mediation.

As you can see, the hardware solution has obvious advantages. However, the problem is its size and portability. The cost of these devices is getting cheaper and cheaper, and it will be affordable for many organizations. But, you can't move such facility beside you. Therefore, software solution is also important issues for overcoming communication barriers.
This shows the framework of possible software solution. This framework targets a situation in which you try to take part in a meeting from remote place using PC, PDA, or mobile phone, or look at pre-recorded video of a meeting of which you could not attend. The screen of these devices is too small to accommodate the whole view of the meeting. Therefore, it is required to select view or create images that can help viewers understand the progress and the content of meeting. However, there has not been good enough technique to carry out such video editing/creation automatically.

Of course, human video editor/engineer can make understandable video just like movie or TV drama scenes. But, it is not practical solution for daily meetings.

To solve this problem, we believe a framework consisting of “communication scene analysis and visual representation” is an effective solution.

The communication scene analysis includes the recognition of human behavior in conversation to understand what’s going on in the conversation. The visual representation issue includes an automatic video editing using the analysis results. For example, switching participant’s view according to flow of conversation.
First let me introduce communication scene analysis.

“Software solution” can overcome communication barriers

- Imagery
- Measuring behaviors
- Recognizing behaviors
- Understand conversations

Visual Representation

- Automatic video editing; e.g. switching participant’s view according to their roles
- Interface design
Communication Scene Analysis

The goal of communication scene analysis is to automatically discover the who, when, where, whom, what, how, and why of our communication. In this study, we focus on who and whom among these interrogatives, and target so-called “who-to-whom” problem. In addition, we focus on two definitions of communication; one is as an act of transmitting/receiving messages between people, and the other is as an act that can affect others' mind, as the result of exchanging messages. First one sheds light on physical aspect of communication, and second one refers to mental/psychological aspect of communication.

We are tackling two problems that correspond to these definitions. One is identifying conversation structure from people's behavior. This allows us to answer the questions about "who is talking to whom" and "who is listening to whom". The second problem is quantifying interpersonal influence in conversation. This can answer questions like "who influenced whom?" and "who was the most influential person?".
Probabilistic Inference of Multiparty-Conversation Structure based on Gaze Patterns, Head Directions, and Utterances

Presented at ICMI 2005

Let's me begin with first problem.
This is the work I have been engaged in for last one and half years.
As I just mentioned, the aim of this study is to identify conversation structure, which can indicate "Who is talking/listening to whom", by observing people's behavior. Here, the conversation structure is defined as the combination of participants and their participation roles. The participation role consists of speaker, addressee, and side-participants who take part in conversation but currently not being addressed. This problem is not so simple as it looks like. Because, when there are more than two people, audio information does not often offer sufficient clues to distinguish the roles.

To identify conversation structure, which can indicate "Who is talking/listening to whom", by observing people’s behavior. Conversation structure is defined as the combination of participants and their roles.

**Participation roles in conversations**

- **Speaker**
- **Addressee**
- **Side-participant** (currently not being addressed)

When more than two persons, the problem is not trivial. Audio information does not offer sufficient clues to distinguish the roles.
Our Approach

- Focus on **gaze behavior** as one nonverbal behavior.
- Hypothesize that gaze behavior can indicate conversation structure
- Develop a probabilistic conversation model
  - Dynamic Bayesian network
  - Participants’ behavior: gaze, head movement, presence/absence of utterance.
  - Bayesian inference with MCMC (Markov chain Monte Carlo)

To solve the problem, we have proposed an approach shown here.

First, we focus on gaze behavior as one nonverbal behavior. Second, we hypothesized that gaze behavior can indicate conversation structure. We have developed a probabilistic conversation model based on a dynamic Bayesian network involving gaze, head movement, and utterances. Using the model, the conversation structure and gaze pattern are estimated by Bayesian inference using MCMC, Markov chain Monte Carlo method.
Nonverbal Behavior

**What kind of messages do we exchange in face-to-face conversation?**

<table>
<thead>
<tr>
<th>Verbal message</th>
<th>30-35%</th>
<th>7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocal language</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nonverbal messages</th>
<th>65-70%</th>
<th>93%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facial expression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eye gaze</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bodily motion, gesticulation</td>
<td>Birdwhistell 1970</td>
<td>Mehrabian 1968</td>
</tr>
<tr>
<td>Paralanguage, prosody</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical contact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpersonal distance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Large part of message is exchanged through nonverbal behavior.

At first, let me briefly mention why I focused on nonverbal behavior, not languages. From past psychological studies suggest that large part of message is exchanged through nonverbal behavior in face-to-face conversation. Examples include facial expression, eye gaze, gesture, and so on.
Among various nonverbal behaviors, it is strongly suggested that eye gaze plays an especially important role. For example, according to Adam Kendon, gaze has three functions such as monitoring others, expression of one's attitudes and intentions, and regulating the flow of conversations. For example, when a speaker ends his utterance, he tends to look at the next speaker as turn-taking cues.

Also, Charles Goodwin indicated that the speaker uses his/her gaze to indicate whom he/she is addressing, and securing their attention is necessary to keep his/her position as the speaker. Hearers need to show their attention by looking at the speaker.

So, this suggests that gaze plays an essential role to organize the conversations in face-to-face setting.

Based on these past findings, [Continue to next page]
Gaze Pattern Analysis

This study analyzed typical gaze patterns in 4-person conversation
(Top 6 relative frequency to chance level, isomorphic-graph category)

Frequent and lasting gaze patterns exhibit unique topologies
- convergence of gaze onto one person
- mutual gaze between two people

Hypothesis: gaze pattern is strong indicator of conversation structure

We tried to find out what kind of gaze behavior actually appears in a conversation. To that end, we manually detect gaze direction of participants by carefully watching video.

Here, let me show you a video sequence of the conversations we are using, and detected gaze patterns.

[Video clip]

In this figure, the thin long line indicates head direction and the short line indicates gaze direction. No gaze line indicates averted gaze status.

Based on these gaze patterns, we calculated the frequency of each gaze pattern that appeared in the conversation. This figure shows six of the most frequent gaze patterns. We found a small number of common patterns. One is the convergence of gaze onto one person, in other words, one person collects the others' gaze more than the others. The second pattern is mutual gaze between two people.

Based on these observations, we hypothesize that gaze pattern is a strong indicator of conversation structure.
And we defined three classes of typical conversation structures, which are indicated by the common gaze patterns. Here we call it "conversation regimes".

First, the regime called "convergence" corresponds to the gaze pattern in which the gazes from many participants converge on one person, (i.e. one person attracts the others' gazes more than the others.) The center person corresponds to the speaker, and the others are addressees. This regime is assumed to indicate monologue.

The second regime, called "dyad link", corresponds to the situation that two people look at each other. During this regime, they exchange messages and can swap the roles of speaker and addressee;

These two people become speaker and addressee, and the others are side-participants. The flow of information is one-to-one, bi-directional. This regime indicates dialogue.

The third regime, called "divergence", corresponds to the gaze patterns that people look in different directions or avert their gaze. In this regime, group conversation does not occur.
We employ Dynamic Bayesian Network.

Conversational regime governs how people interact
i.e. Regime state controls dynamics of gaze patterns & utterances

Based on the hypothesis that links gaze pattern to conversation structure, we built a conversation model based on dynamic Bayesian Network. This model assumes that the state of conversation regime changes over time according to Markov process, and it controls dynamics of gaze patterns and utterances. Also, we assume that gaze direction can be inferred by head direction, because direct measurement of gaze direction is difficult in natural conversation scene.
Based on the model proposed, we formulated the problem as the inference of the joint posterior distribution of all unknown variables for the given observable variables like this, following the Bayesian framework. According to Bayes' theorem, the posterior distribution is proportional to the product of likelihood function and prior distribution. However, due to model complexity, the analytical solution is intractable to calculate, so we decided to employ the approximation method, a MCMC (Markov chain Monte Carlo) method known as the Gibbs sampler.

Let me omit the detail algorithm due to time shortage.
Next, let me move on to the experiment. We targeted 4-person group conversations. The participants were 4 women within the same age bracket. They were instructed to hold a discussion on a given topic and try to reach a conclusion as a group within five minutes.

The topics were chosen from various hard and soft topics such as "Are love and marriage the same or different?" and "Should the death penalty be legitimate or not?"

The participants were seated like this. The head directions were measured at 30[Hz] using POLHEMUS magnetic-based (6-DOF) sensors, which were attached to their heads with hair bands. Audio data were recorded by clip-on microphones attached to each participant, and utterance intervals were manually extracted using a waveform editor. Also, video sequences, (whole shot and bust shots) were recorded at 30[frame/sec]. (These data were synchronized at the unit time step of 1/30[sec].)

Let me show you what the measured data look like.

[Video Clip]

Here is part of recorded video and measured head directions. This shows a temporal series of head directions, azimuth angle against the X axis, and this shows the presence and absence of utterances at each time step. These two were input to the Gibbs sampler.

This shows the sensor output for each participant.
This shows the measurement setting. A receiver for measuring head-direction was attached to each person, and the transmitter was hung from the ceiling. A number of cameras, about 10 to 12, were used as shown here.
Experimental Result: Gaze

Confirmed reasonable accuracy in estimating gaze direction

Next, let me show you some of the results.

First, this shows the time series of gaze direction of each participant. The red line shows the estimates and the blue line shows the ground truth, which was manually created by watching the video sequences. This Table shows the average correct ratio of the number of frames wherein estimates and ground truth coincided.

Let me show you one part of the results on video.

[Video Clip]

The longest thin line indicates head direction, and the middle thin line indicates the ground truth. The widest line indicates the estimated gaze direction. Also, a person making an utterance is indicated by a small green circle.

As you can see in this graph, most errors (65 to 78%) were related to the averted gaze status. This is because humans can avert/turn their gaze from/on someone without moving their head, e.g. using a sidelong glance.

Despite such intrinsic limitations of using head direction, reasonable accuracy in estimating gaze direction was confirmed.
Result: Identified Regimes

Experiments show regime sequence can track flow of conversation

Next, let me show you a video that demonstrates regime estimation.

[Video Clip]

The single circle represents the center person in the case of regime convergence, while the pair in the dyad link is represented by the two circles.

No circle indicates regime divergence.

At first, person 4 talked to all others, and others listened to person 4. This form of communication was indicated by the estimated regime, convergence to person 4.

Next, person 2 returned strong response to person 4 and person 4 looked at person 2 to confirm her response. There was mutual gaze between person 2 and person 4, and the corresponding regime was dyad link between person 2 and person 4.

Next, person 2 kept on speaking and person 4 offered the floor to person 2. At the same time, person 3 turned her gaze from person 4 to person 2 to hear what person 2 would say.

Finally, person 2 talked to person 1 and person 1 replied. This is suggested by the dyad link between person 2 and 1.

As you can see, this flow of conversation can be tracked as a sequence of regimes, and this indicates the effectiveness of our proposed framework toward the automatic identification of conversation structure.
Next, let me explain how we evaluated regime estimation. This method based on manual annotation; each utterance interval is tagged with its class and its directionality.

For example, person 1 says "I'm looking for a train station. Could you give me directions?" followed by "Sure, Which station?" from person 2. The label for person 1's utterance includes expressing opinion (giving information) toward person 2, asking P2 a question, and the listener is P2. The label for P2 consists of responding to P1's question, expressing her opinion toward P1, asking a question and the listener is P1.

Here, let me show you some labels extracted from a video.

[Video Clip]
Accuracy of regime estimates

Accuracy is calculated by matching the regime estimates and utterance labels.

Decision rules:

- **Regime convergence**: \( R_i^C \)
  \[
  \left\{ \begin{array}{l}
  \text{-Person } P_i \text{ talked to all others} \quad \text{OR} \\
  \text{-Person } P_i \text{ talked to a person, and all people listened to } P_i.
  \end{array} \right.
  \]

- **Regime dyad-link**: \( R_{(i,j)}^{DL} \)
  \[
  \left\{ \begin{array}{l}
  \text{-Person } P_i \text{ questioned/responded to only } P_j \quad \text{OR} \\
  \text{- } P_i \text{ talked to only } P_j \text{ and only } P_j \text{ listened to } P_i \text{ OR } P_i \leftrightarrow P_j
  \end{array} \right.
  \]

Results: Ave. 81.9% (Conv. 85.0%, Dyad-link 77.7%, Div. 53.2%)

Based on this annotation, the accuracy of conversation regimes is calculated by matching the regime estimates and utterance labels.

Examples of decision rules are shown here. For regime convergence, person \( P_i \) talked to all others, or person \( P_i \) talked to a person, but all people listened to \( P_i \).

For regime dyad-link. Person \( P_i \) questioned or responded to only \( P_j \). Or, \( P_i \) talked to only \( P_j \) and only \( P_j \) listened to \( P_i \).

The results for a 5 minute segment are shown here. The average accuracy was 81.9%. This indicates that our method can outperform methods based on just a single modality, say gaze or utterance.
The future works include the following topics: Dealing with more realistic situations including the presence of furniture and materials, exploring social / psychological aspects using other nonverbal behavior, and developing sensing methods like visual head tracking. Here, let me briefly show you the head tracking, we are developing now, by video.
Communication Scene Analysis

Goal: Discover who, when, where, whom, what, how, and why of our communication

We focus on two definitions of communication

- as act of transmitting/receiving messages
- as act that can affect others’ mind

And focus on "Who-To-Whom Problems"

1) Identifying Conversation Structure from Behavior
   Who is talking/listening to whom?

2) Quantifying Interpersonal Influence in Conversation
   Who influenced whom?

Next topic is quantifying interpersonal influence.
Quantifying Interpersonal Influence in Face-to-Face Conversation Based on Visual Attention Patterns

Presented at CHI 2006 Work-In-Progress Session.

This work was just presented at CHI held in Montreal this week, and currently in-progress.
Attention-based Influence

**Aim:** Propose measures that can quantify person-to-person influence in face-to-face conversation

**Concept:**
*The more you speak, the more you can affect others, only if your speech is acknowledged / attended by others.*

**Idea of attention-based influence:**
- Amount of influence
- Length of speech
- Amount of Visual Attention

Based on estimated conversation structures and gaze patterns.

The aim is to develop a method for automatically quantifying interpersonal influence in face-to-face conversation, which should provide answer “who influenced whom?” and “who was the most influential?”.

The basic concept of our influence measures is that the more you speak, the more you can affect others’ mind, only if your speech is acknowledged or attended by others. In general, the first line would be considered as true statement, but we believe the second line is also important factor to measure the influence of people.

Based on the concept, we devised the idea of attention-based influence, as shown here. The amount of influence is proportional to the length of speech, that is weighted by the amount of visual attention paid to the speaker. These are obtained as the estimated conversation structures and gaze patterns.
Calculating Influence

**Bilateral Influence**

\[ I(i, j) := I_M(i, j) + I_D(i, j) \]

- Influence from monologue: Amount of attention paid to speaker in monologue
- Influence from dialogue: Duration of dialogue involving persons i and j

**Outgoing Influence**

\[ I_{OUT}(i) := \sum_{j=1}^{N} I(i, j) \]

Amount of influence that person i had on others

**Incoming Influence**

\[ I_{IN}(i) := \sum_{j=1}^{N} I(j, i) \]

Amount of influence that person i received from others

Using the estimated sequences of the conversation structure and gaze patterns, that I explained earlier, some influence-related measures can be calculated as shown here.

The most basic one is called "bilateral influence". It indicates the amount of influence that one person has on another. It is defined as summation of influence achieved in monologues and that achieved in dialogues. The influence achieved in a monologue is defined as the amount of gaze directed to the speaker in the monologue. That achieved in a dialogue is based on the duration of dialogue involving two people.

Based on bilateral influence, some measure, which can represent personal characteristics, can be defined that covers outgoing influence and incoming influence. Outgoing influence is the amount of influence that one person had on the others, and incoming influence is the amount of influence that the person received from the others. These are calculated by summing the bilateral influences.
Personal Characteristics

We have developed two measures to represent personal characteristics in conversation; one is influence balance and the other is participation level, based on outgoing and incoming influence. The influence balance indicates the main role of each participant, speaker or listener. It is calculated by subtracting incoming influence from outgoing influence. The participation level can indicate the level of activity. It is calculated by adding incoming and outgoing influence. These measures can be displayed as shown here; you can clearly see the role and position of each participant.

Influence Balance
indicates main role of participants

\[ \Delta I(i) := I_{OUT}(i) - I_{IN}(i) \]

+ acted as speaker
balanced
- acted as listener

If outgoing influence of a person is greater than incoming influence, he/she mainly acted as speaker.

Participation Level
indicates level of activeness and involvement.

\[ I(i) := I_{OUT}(i) + I_{IN}(i) \]

highly active speaker
balanced participants
mainly listener

Participation Level
indicates level of activeness and involvement.
These measures can be visualized as the influence network shown here. Person-to-person influence is indicated by the thickness of the arrows. A large node indicates large outgoing influence. So, you can easily see who the most influential person in the conversation was.

For example, person 3 was the most influential and person 2 was the least influential person. Also, person 2 was greatly influenced by person 3, but person 2 did not influence person 3.
Conversational Characteristics

Interactivity

\[ I_{D/M} := \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} I_D(i, j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} I_M(i, j)} \]

large: chat-like interactive conversation
small: lecture-like conversation

Centralization

\[ I_C := \frac{\sum_{i=1}^{N} \max_{j} I_{OUT}(j) - I_{OUT}(i)}{(N-1)^2} \]

- \( I_C \approx 1 \): one person dominates others
- \( I_C \approx 0 \): everyone has equal influence

From four different conversations:

- highly interactive and a dominant speaker exists
- less interactive and no dominant speaker exists

Also, two measures that can characterize entire conversations can be obtained. One is interactivity, which can be calculated as the ratio between influence from monologue and that from dialogue. A large interactivity indicates the conversation was chat-like interaction while small interactivity indicates a lecture-like conversation.

Finally, influential centralization indicates the dominancy of conversation. I borrowed this definition and notation from social network analysis. If one person dominated the others, that is, only one person speaks and attracts attentions, the value becomes 1. On the other hand, if everyone had equal influence, this becomes zero.

These measures can be visualized as in graph. Four different conversations are depicted in this graph. It clearly shows the conversational characteristics.
Discussion

Can proposed measure of “influence” reflect actual “influence in conversation”?

Three levels of influence

- [Highest Level]: Belief / opinion (mental state)
- [Middle Level]: Conversational behavior (listening and speaking)
- [Lowest Level]: Physical looking behavior (paying attention)

The amount of “influence” is not necessarily proportional to the number of times, length of speech, and amount of attention.

- Sometimes, one word is worth 100 sentences.

This is where we are now. Here, the question that must be answered is that “can proposed measure reflect actual influence in conversation?”.

To answer that question, we have to think about the three levels of influence; the lowest level is the influence on physical gaze behavior. The middle level is the influence on conversational behavior like speaking and listening. The highest level is the influence on one’s beliefs or opinions, i.e. mental state. We believe that our measures can deal with the first two levels.

But, it is difficult to deal with the highest level.

This is because that the amount of influence is not necessarily proportional to the number of times, length of speech, and the amount of attention. Sometimes, one word could be worth 100 sentences, depending on context of conversation and position/rank of a person in an organization.
Toward “Software” Solution

Communication Scene Analysis
- Imagery
- Measuring behaviors
- Recognizing behaviors
- Understand conversations

Visual Representation
- Automatic video editing; e.g. switching participant’s view according to their roles
- Interface design

“Software solution” can overcome communication barriers

Let me move to the final topic; visual representation.
Its title is "Automatic Video Editing Based on Participants’ Gaze in Multiparty Conversation".

This work was mainly conducted by Yoshinao Takemae and I was his supervisor.
The goal of this method is to automatically create video clips that can clearly convey to viewers the conversation structures and addressees' responses.

To that purpose, we have developed a novel video editing rule, which is a switching multiple cameras that separately capture each participant. The rule is quite simple; selecting the shot of person who is attracting others' gaze more than anyone else. In short, the focus of attention is presented to the viewers as shown this figure.

Goal and Approach

**Goal:**
Automatically create video that can clearly convey to viewers
- Conversation structures; who is talking to whom
- Addressees' responses

**Approach:**
Developed new video editing rules, (Switching multiple cameras that capture participant's shot)

Selecting the shot of person who attract others' gaze more than anyone else

*Focus of attention is presented to viewers!!*
Basis of Method

(1) Consider problem with current voice-switching scheme
   It selects only speaker's shot.
   So, viewer can not understand the addressees and their response.
   This must be reason why viewers feel it difficult to understand conversations from recorded video.

(2) Observe actual participants’ gaze behavior
   ▪ To understand and participate in the conversation, participants constantly decide where to look at.
   ▪ (Side-)participants tend to look at speaker and addressee alternately.

(3) Develop video-switching rules based on gaze behavior
   Assumption: The more people look at someone, the more important information must exist there for both participants and viewers.

Although the editing rule is simple, it has sound basis. I believe.

Before starting the study, we considered the problem with current video editing method, which mainly depends on voice-switching, which selects speaker's shot time to time. So, viewer feels it difficult to understand who the addressee is, and their responses.

Second, we observed the gaze behavior of actual participants in conversations. Since participants constantly decide where to look at, in order to understand and participate in the conversation, we thought that their looking behavior would provide good hint. From observation, we have found that side-participants tend to look at speaker and addressees alternately.

Finally, we have develop the video-switching rules based on the gaze behavior. The assumption is that the more people look at someone, the more important information must exist there for both participants and video viewers.
Demonstration

3-person conversation; person 3 is trying to persuade person 1.

Discussion topic: “Is it necessary to rank people?”

Let me show you one example. The target is three-woman conversation, and three cameras capture their separate images as shown here. The video clips I will show you include a discussion scene. They are discussing on a topic, “Is it necessary to rank people?” and person 3 is trying to persuade person 1.
Demonstration; voice-switching

In this video clip, person 3 is trying to persuade person 1, that is, person 1 is addressed by person 3. However, as you might notice, video does not include person 1’s shot, because she did not say anything. Therefore, voice-switching scheme can not convey person 1’s response. This is a typical drawback of voice-based switching.

Person 3 is trying to persuade person 1 (= person 1 is addressed by person 3). But video does not include person 1’s shot, because she did not say anything. Therefore, voice-switching scheme can not convey person 1’s response.
Next, let me show you our result of gaze-based switching.

[Video Clip]

This clip includes person 1’s shot and you can clearly see her response. She did not say anything, but clearly expressed her disagreement with the opinion of person 3. As you can see in this example, silent response, i.e., nonverbal behavior is an important factor in understanding the conversation.

Gaze-based switching clearly convey person 1’s response; she expressed her disagreement with the opinion of P3. Silence response expressed by nonverbal behavior is also important in understanding conversation.
Conclusion and Future Works

Conclusions

- Conducted several subjective evaluations; comparison with other video editing schemes
- Our method outperforms other schemes; viewer more accurately understand addressees and their responses

Future works

- Evaluating viewers’ understanding in terms of content of conversation
- Exploring more effective visual representation schemes for various devices
- Incorporating conversation structures and interpersonal influence into editing rules

Let me conclude this topic.

So far, we have conducted several subjective evaluations including comparison with other visual representation schemes. And we have confirmed that our method outperforms other assessed, and that viewers more accurately understand addressees and their response.

Future works includes evaluating viewers’ understanding in terms of content of a conversation, and exploring more effective visual representation schemes for various devices, and incorporating conversation structures and interpersonal influence into editing rules.
Conclusions of today’s talk

- t-Room; immersive teleconferencing system
- Conversation Scene Analysis
  - Inferring Conversation Structure based on Gaze Behavior
  - Quantifying Interpersonal Influence in Conversation
- Automatic video editing based on focus of attention

Finally, let me conclude my talk. Today, I have introduced our research activities, t-room, conversation scene analysis, and automatic video editing.

This concludes my talk.

Thank you for your attention.
References

K. Otsuka, Y. Takemae, J. Yamato, and H. Murase,
“A Probabilistic Inference of Multiparty-Conversation Structure based on Markov-Switching Models of Gaze Patterns, Head Directions, and Utterances,”

K. Otsuka, J. Yamato, Y. Takemae, and H. Murase,
“Quantifying Interpersonal Influence in Face-to-face Conversations based on Visual Attention Patterns,”
In Proc. CHI2006 (To appear)

K. Otsuka, J. Yamato, Y. Takemae, and H. Murase,
“Conversation Scene Analysis with Dynamic Bayesian Network Based on Visual Head Tracking,”
In Proc. ICME2006 (To appear)