ABSTRACT
A novel system, called MM+Space, is presented for recreating multiparty face-to-face conversation scenes in the real world. It aims to display and playback pre-recorded conversations as if the people were talking in front of the viewer(s). This system consists of multiple projectors and transparent screens, which display the life-size faces of people. The key idea is the physical augmentation of human head motions, i.e. the screen pose is dynamically controlled to emulate the head motions, for boosting the viewers' perception of nonverbal behaviors and interactions. In particular, MM+Space newly introduces 2-Degree-of-Freedom (DoF) translations, in forward-backward and right-left directions, in addition to 2-DoF head rotations (nodding and shaking), which were proposed in our former MM-Space system. The full 4-DoF kinetic display is expected to enhance the expressibility of head and body motions, and to create more realistic representation of interacting people. Experiments showed that the proposed system with 4-DoF motions outperformed the rotation-only system in the increased perception of people's presence and in expressing their postures. In addition, it was reported that the proposed system allowed the viewers to experience rich emotional expressibility, immersion in conversations, and potential behavioral/emotional contagion.

Categories and Subject Descriptors
H1.2 [Models and Principles]: User/Machine System — Human Information Processing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMI ’13, December 9–12, 2013, Sydney, Australia
Copyright 2013 ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522854.

MM+Space: $n \times 4$ Degree-Of-Freedom Kinetic Display for Recreating Multiparty Conversation Spaces

Kazuhiro Otsuka
NTT CSL
3-1 Morinosato Wakamiya
Atsugi-shi, Japan
otsuka.kazuhiro@lab.ntt.co.jp

Shiro Kumano
NTT CSL
3-1 Morinosato Wakamiya
Atsugi-shi, Japan
shiro.kumano@lab.ntt.co.jp

Ryo Ishii
NTT CSL
3-1 Morinosato Wakamiya
Atsugi-shi, Japan
ishi.ryo@lab.ntt.co.jp

Maja Zbogar
NTT CSL
3-1 Morinosato Wakamiya
Atsugi-shi, Japan

Junji Yamato
NTT CSL
3-1 Morinosato Wakamiya
Atsugi-shi, Japan

Keywords
face-to-face conversation, telepresence, projection mapping, multimodal interaction, computer-mediated communication

1. INTRODUCTION
Face-to-face conversation is one of the most basic forms of communication in daily life and group meetings are used for conveying/sharing information, understanding others' intention/emotion, and making decisions. Visual communication research aims to realize smooth communications among spatially separated places, as natural as is possible in real face-to-face settings. Toward designing better communication systems, the authors have focused on the essential sub-problem of how people perceive and understand others' conversations, and the mechanism behind it.

In face-to-face conversations, people exchange not only verbal messages, but also nonverbal messages, such eye-gaze, head and body gestures, facial expression, and prosody [1]. Therefore, it is essential to reproduce these modalities in the remote environment. Among the various nonverbal behaviors used, head motion plays an important role, especially as an indicator of visual attention; it provides essential functions such as monitoring others, expressing one’s attitudes/intentions, and regulating conversation flow [8]. Head gestures such as nodding, shaking, tilt, also play important roles in conversations [13], e.g. a sign of addressing/questioning by speakers, and back-channel response by hearers.

Considering the importance of nonverbal information, the authors proposed a novel idea and introduced the concept system called MM+Space, for recreating multiparty conversation scenes in the real world [18, 17]. The key idea is the physical augmentation of human head motion, i.e. the screens that display the face of each meeting participant, are dynamically controlled to recreate the actual human head motions. This system consists of multiple projectors and transparent screens, which display the life-size faces of people, and are attached to actuators that provide 2-Degree-of-Freedom (hereafter DoF) rotation, in nodding and shaking directions. Some experiments have confirmed its effectiveness in boosting the human perception of people's gaze di-
reactions and understanding of conversation structures, such as “who is talking to whom”.

This paper proposes a novel system, called MM+Space ¹ as an extension of the former MM-Space. MM+Space introduces two new DoF translations, forward-backward and right-left directions, in addition to the 2-DoF head rotations (noodling and shaking directions), as offered by our MM-Space system. We call this new device the n×4-DoF Kinetic Display, where n denotes the number of conversation participants. The 2-DoF translations are generated with XY-stages, which consist of two orthogonal linear sliders driven by linear motors.

The MM+Space is expected to enhance the expressibility of head and body motions, and to create more realistic social telepresence of interacting people. First, the forward-backward position of the screen represents leaning forward/backward postures, which can reveal one’s attitude and social distance toward others, such as the level of interest and intimacy. Second, combined translations and rotations can enhance the dynamic range of head gestures, which are expected to enrich emotional expressions, e.g., deep slow nodding indicates deliberate thought/understanding of what others said, a rapid retreat motion can express the surprise, and coordinated head nodding among people indicates interpersonal empathy. Furthermore, unconscious body movements can be taken as a sign of human presence.

The idea of head-motion augmentation comes from biological motion [7, 16, 3] and mental attribution [5]. Humans tend to anthropomorphize lifeless objects by assigning social contexts to movements, even if the objects are simple geometric shapes such as points and squares. Based on this, we hypothesize that realistic kinematics offers strong cues for better understanding human communications; for this, we employ flat square screens for both displaying faces and physical motions, as the most simple and primitive form of the screen. In addition, physical motion is also effective for intuitively representing the entire conversation space at a time, because human peripheral vision has high sensitivity to motion cues [10, 23], e.g., when one is paying attention to a speaker in his/her central visual field, one can also notice nodding by others in one’s peripheral view. Authors believe that this can contribute to establishing the realistic virtual experience of remote meetings.

The ideas and concept above make our approach unique and distinct from other related works, including projection-based telepresence avatars, such as those that use spherical screens (SphereAvatar [19]), cylindrical screens (TeleHuman [9]), and face-shaped screens (TalkingHeads [15]). In addition, the spherical and cylinder screen are significant different in shape from real human faces and geometrical shapes. This gives the viewers un-natural impressions. Therefore, we believe that the flat screen is the optimal choice for our system. In addition, our approach is closely related to telepresence robots or robot avatars [24], which aim to create realistic physical entities that can communicate with humans. However, many robotics approaches have difficulty in reproducing natural motions, which results in giving unnatural impressions to users, because human can easily notice stiff robot-like motion. Furthermore, a realistic android suffers from the Uncanny Valley problem [14]. Unlike the usual robotics approach, our approach makes mechanical motion just a trigger for the user’s perception, and the user can fully recognize the human motion in the images. So even if the range of physical motions is limited, our system can easily ensure the naturalness of the expressions displayed.

The concept of spatial distributed displays, employed in our MMSpace series, which assigns each meeting participant to a separate display, can be found in earlier works such as Hydra [21]. The aim was to keep the spatial consistency among places and to allow users to perceive who-to-whom cues in a clearer way. This idea has been extended to immersive teleconference environments, surrounded by wall-like panel displays such as t-Room [6]. However, the static distributed displays is likely to raise a problem with the users’ perception, e.g., when a user is straight looking at a person, it is hard for him/her to notice what others are doing, because the human visual field can not cover all people at a time, and the users often need to look around. This limitation triggered our insight and led to our kinetic displays.

Note that MM+Space integrates flat panel loudspeakers with the transparent projection screens and so gives an almost perfect match between visual face position and sound source of human voice, i.e., the viewers can perceive that the human voice comes straight from the displayed human face, whereas MM-Space places loudspeakers in front of the screen on the table, which creates a noticeable audio/visual inconsistency.

Experiments involving naïve viewers were conducted to compare MM+Space with the MM-Space and the static condition without any motion. Paired comparison among the three conditions showed that MM+Space with 4-DoF motions outperformed the rotation-only MM-Space system in terms of the increased perception of people’s presence and in expressing their postures. In addition, it was reported that viewers experienced rich emotional expressibility, immersion into conversations, and possible behavioral/emotional contagion.

This paper is organized as follows. Section 2 describes the proposed system. Section 3 details the experiment and discusses the results. Finally, our conclusion and some discussions are presented in Section 4.

2. PROPOSED SYSTEM: MM+SPACE

2.1 Overview

Fig. 1 illustrates the proposed MM+Space system from the viewer’s viewpoint. Fig. 2(a) shows its original scene, and Fig. 2(b) provides a block diagram of the system. As shown in Fig. 1, MM+Space consists of multiple screens, each displays the face of a different meeting participant. They are located to recreate the spatial configuration of the participants, as seen in Fig. 2(a). Each participant’s face is displayed on a flat transparent screen whose position and pose are dynamically changed in sync with the head motion of the participant.

In the block diagram in Fig. 2(b), input parts consists of multiple cameras for capturing participant’s faces, microphones, and sensors that can measure the 6-DoF status of each participant’s head, i.e. 3-D position and 3-DoF rotation angles. As in MM-Space, each camera was set to capture

---

¹In MM+Space or MMSpace-Plus, “MM” takes multiple meaning including multimodal, multiparty, meeting, multilocation, and minimal design. “+” symbolizes the orthogonal axes of translations.
Figure 1: Overview of MM+Space from viewer’s point of view

Figure 2: (a) Original scene, (b) Block diagram of MM+Space

each person’s face from (roughly) the full frontal direction, and lapel microphones were attached to each person.

The processing parts provide visual face tracking, background removal, image mapping, and control signal generation. The output parts consist of projectors, Pan-Tilt Units, XY stages, and Giant Magnetostrictive Material (hereafter GMM) actuators. The Pan-Tilt Unit, in short PTU, controls 2-DoF rotational screen poses of the human head in nodding and shaking directions. Note Pan motion of PTU corresponds to human shaking direction, and Tilt corresponds to the nodding direction. Note here the human tilting action refers to cocking one’s head on one side, and its mechanical replication currently exceeds our scope. The XY stage controls the 2-D horizontal position of the screen panel. The XY-stages consist of two orthogonal linear sliders. Here linear motors were used to drive the sliders, because of their silent operation. The GMM actuator is a vibrator that drives the screen panel as the loudspeaker.

2.2 4-DoF Kinetic Display

Fig. 3(a) shows a frontal view of the screen. Each screen is highly transparent but includes a diffusive material that catches the projector’s output and makes it visible to the viewer. Each screen has its own LCD projector behind it. Fig. 3(b) shows the backside view of the screen and the mechanics for PTU, XY stage, and GMM actuators. Looking the equipment from top to bottom, the screen is supported by a panel holder, which is connected to the head of the PTU. The PTU is set on the XY stage fixed on the top of a tripod. The GMM actuator is supported by the panel holder, and has a point contact on the screen panel, with a constant pressure given by a spring suspension system. The panel holder offers two-point screen support, which affords room for audible oscillations triggered by the GMM actuator placed at the center of the holder. This panel holder, specially designed for MM+Space, is made of aluminum, and manufactured by CNC machining.

The GMM actuator is based on the phenomenon called Magnetostriction. It is the property of ferromagnetic materials that causes them to change their shape or dimensions under external magnetic fields [4]. Giant magnetostrictive materials are functional materials that offer very large magnetostrain and high energy density. We employed a GMM device to realize the flat panel loudspeaker, because it can be fully integrated with the screen panels without negative optical effects.

2.3 Image Processing

In the processing part, visual face tracking [12] measures the head position and pose of the meeting participants on the camera images. Background removal creates images that focus on the participants [2, 20]. From the background-removed images, the face region is extracted based on the face position measured by the face tracker [12]. Image map-
ping generates the projection images that are skew-free, depending on the pose and position of screen, which are controlled with PTUs and XY stages. Image processing part is the same as in MM-Space [18, 17].

2.4 Measuring face position/pose

For controlling PTUs and XY stages, the face position and pose are measured from the 6-DoF output of the sensor, which is currently attached to the head of each person. This is because 2-D image-based tracking, used for determining 2-D face position, fails to yield accurate or reliable 3-D position, especially in the depth direction toward the camera, for controlling translational XY stages. A future implementation might include the use of depth cameras like Microsoft’s Kinect.

2.5 Controlling PTUs

The process used for generating PTU control signals follows that used in MM-Space [18, 17]. The PTU must control screen movement to emulate head rotation as closely as possible. The speed and torque attributes of the PTU are limited, and the range of screen rotation must be restricted so that the viewer can clearly see the face on the screen. From this perspective, the PTU control signal is generated by reshaping the time series of head rotation angles, including de-noising, temporal downsampling, and down-scaling amplitude. Here, the scaling factors for pan and tilt motions were set to 0.4 and 0.1, respectively.

2.6 Controlling XY stages

The XY coordinates of measured head position on the horizontal plane are used to control the XY stages. Fig. 4(a) shows an example of the trajectory of XY coordinates over time. Because the sensor is attached at ear level on the back the person’s head, the measured 3-D position has a offset relative to the center of the head.

As seen in Fig. 4(a), the trajectory exhibits a circular arc drawn by rotating head with a sensor in shaking directions. Here, the XY values of the sensor output is corrected with regard to the rotation center when head shaking. Without the correction, the resulting motions would be unnatural; body position also changes as head shakes. Because the exact relative position between the sensor and the rotation center is not available, the approximate relation is calculated by minimizing the variance of the estimated rotation center when the person widely turned his/her head. Fig. 4(b) shows an example of the XY trajectory after correction. As seen in Fig. 4(b), the arc-shaped trajectory has disappeared and XY positions now seem to be invariant to head rotation in the shaking direction. Also, the most of the trajectory now lies inside the travel limits, the square framed region in Fig. 4. Note there remain translational components that accompany head rotations in nodding and tilting directions, and the coordinated motions of PTU and XY stages can visibly enhance the head nodding/tilting gestures.

Next, the denoising and smoothing process are applied to the time series of XY positions, because they include significant amount of measurement noise, as shown in Fig. 4(a) and 4(b). The XY stage control signal is then generated by
reshaping the time series of XY position, in a manner similar that used for PTU control. This involves the thresholding of velocity, making small velocities zero, and bounding the position to within the travel limits of the screen. Fig. 4(c) shows the XY trajectory of the control signal; it is a smooth continuous trajectory without noise. Unlike the PTU control signal, no down-scaling is used, i.e. the screen panel translates at the same speed as the original head motion, the exception being the constraint of the travel range. If the head exceeds the range, the screen panel stays at the border until the head returns to the range.

Fig. 4(d) shows an example of the actual movement trajectory of the XY stage, as indicated by the encoder feedback from the XY stage. Comparing Fig. 4(c) and 4(d), you can see that the XY stage well replicated the input control signal.

2.7 Example scenes

Fig. 5 show some snapshots of characteristic scenes 2. Fig. 5(a) shows a questioning scene. The speaker leaned forward to the addressee and nodded to her. The behaviors were reproduced by the panel moving straight toward the addressee with tilt down at the end of the move. Fig. 5(b) shows a hearer’s back-channel response. Her slow nodding was mechanically recreated with coordinated rotation and translation. Because the PTU’s tilt motion (in human nodding direction) produces significant mechanical noise, the angle range must be small (e.g. 1/10 of real amplitude). Fortunately, the added translations improved the reality and naturalness of the behavior. Fig. 5(c) shows the reaction of a hearer who started back in surprise with laughter. This was expressed by a rapid retract motion with a small twist of the panel. Because highly emotional responses like this involves body motions, the proposed system is thought to be an effective way of expressing them. In addition, small continuous fluctuation confirms the presence of the person.

3. EXPERIMENTS

Experiments were conducted to compare the impressions of viewers in three different display conditions, \( \text{MM+Space} \) with full 4-DoF motions (called the XY condition), \( \text{MM-Space} \) with only rotations (called PTU condition), and static condition without any physical motions, and to try to discover the characteristics specific to the proposed \( \text{MM+Space} \).

3.1 Experimental Settings

Digital cameras (Point Grey Research’s Dragonfly™) captured VGA images at 30 [fps]. The PTUs were Directed perception’s PTU-D46-70. Maximum angle velocity is 60[deg/sec]. The screen size was 415×415[mm]. The transparency was 97%. The projector was an EPSON 1925W (4000[lm], XGA resolution). The XY stage was Aerotech Inc. ANT130-110-X/Y. The travel range is 110×110[mm], and maximum velocity is 350[mm/sec] on each axis. The head position/pose was measured with Polhemus FASTRAK™ at 30[Hz]. The GMM actuator was Jigbo Inc. JB-GM02.

3.2 Experiment 1

To gather preliminary information, Experiment 1 compares the difference in the viewers’ instantaneous impressions, and so precludes full contextual information. To achieve this we used short clips to focus on the impact of physical screen motion as a nonverbal cue. Here we used the paired comparison method based on Thurstone’s Case V model with Maximum Likelihood Estimates [11, 25].

Data

In the experiments in this paper, we employed 2 group 4-female conversations, and used 4 conversion sessions, two from group 1, denoted C1-1 and C1-2, and two from group 2 from group 2, denoted C2-1 and C2-2. The conversation participants were instructed to hold a discussion on a given topic and try to reach a conclusion as a group within 5 minutes. The discussion topic of C1-1 and C2-1 was “Is love and marriage same or different?”. The topic of C1-2 was “Should euthanasia be legalized?”, and that of C2-2 was “Should tax privilege be given to full-time housewives?”. C1-1, C1-2, C2-1, and C2-2 had durations of 7.7, 5.2, 5.1 and 5.5 minutes, respectively. Here, for the paired comparison, 7 scenes (short clips) in total were extracted from these conversation sessions. Each clip was 10 seconds long. The number of clips, 7, was chosen by balancing data amount against viewer tiredness (about 21 min. in total). The temporal intervals of the clips was selected based on saliency measures in terms of the amounts of facial expressions, its change over time, head gestures, gaze direction changes, utterance, silence, and interpersonal empathy. This saliency

\[ \text{Figure 5: Example scenes, (a)questioning, (b)hearer’s back-channel response, (c)emotional reaction} \]
measure-based selection (with highest level criteria) is used to choose typical scenes from the conversations as regards multimodal nonverbal cues exchanged among people.

**Subjects**

We employed 9 subjects who were paid and who did not participate in the conversations, nor in the past experiments held in [18, 17]. They were females in their 20’s ∼ 30’s, close in age to the conversation participants. Hereafter, we also call the subjects the viewers.

**Procedure**

Each viewer participated in a series of trials. In each trial, the viewer watched a single clip in two difference conditions, chosen from the three conditions, one after the other. Total number of trials for each viewer was 21, i.e. 7 scenes multiplied by 3 pairs of three condition combinations. The order of viewing scenes and condition was randomly mixed.

**Questionnaire**

After viewing each clip pair, the viewer was asked to fill in a questionnaire at the end of the clips that asked them choose one of two conditions based on her impression. More specifically, they were asked 8 questions such as “Which condition did you feel more (clearly) the a)naturalness, b)atmosphere, c)presence of people, d)emotion, e)gestures, f)posture, g)gaze direction, and h)enjoyability?” Time constraint for filling in the questionnaire was 30 seconds for each trial.

**Result and Discussion**

Fig. 6 shows the Thurstone’s scale of three conditions, for each of 8 properties, a)naturalness, b)atmosphere, c)presence of people, d)emotion, e)gestures, f)posture, g)gaze direction, and h)enjoyability, based on the total n = 63 sample data (9 viewers watched 7 scenes). A greater value indicates that the property was more strongly felt. In addition, in Fig. 6, the bracket bridging two conditions indicates that the condition with greater value was statistically significant at the 5% level, than the condition with lesser value; the p-value is on the bracket. The null hypothesis was that both conditions had the same rate (i.e. 50% each) of viewer choice, while the alternative hypothesis was that one condition was chosen by viewers at a higher rate than the other condition. For this multiple comparison, Shaffer’s method [22] was used to adjust the significance levels.

As you can see, in all properties, the static condition was separately located to the left of XY and PTU conditions. This indicates that our approach, the physically augmented display of head motions, boosted the viewers’ perception/ impression as regards the properties. For almost all properties, XY and PTU conditions exhibited statistical significance over the Static condition, the exception was the case of XY condition in a)naturalness.

Comparing XY and PTU conditions, XY condition outperformed PTU condition in (c)presence and (f)posture, with statistical significance of 5%. This indicates that the additional translational motion can increase the sense of social presence, and can express the participant’s postures more than is possible with just rotation. In other properties, no statistical significance was observed. Both conditions showed about the same Thurstone scale in terms of (b) atmosphere, (d)emotion, and (e)gesture. In (h)enjoyability, XY condition was located at a higher position than PTU; the difference almost reached the 5% significance level (p = 0.051).

In gaze direction, PTU condition yielded higher values than XY (p = 0.13). This could imply that PTU condition is able to more clearly express the gazing direction by its simple head rotation, whereas the higher DoF motions provided by the XY condition could entail some ambiguities in the interpretation by the viewers.

![Figure 6: Result of paired comparison in Thurstone’s scaling. a)naturalness, b)atmosphere, c)presence of people, d)emotion, e)gestures, f)posture, g)gaze direction, and h)enjoyability](image)

**3.3 Experiment 2**

Experiment 2 aims to characterize the proposed method, MM+Space, by extracting more detail comments from the viewers.

**Data and Subjects**

Two conversation sessions, C1-1 and C2-1, were used in this experiment. The subjects were those from Experiment 1.
Procedure
Each subject separately watched the two conversation sessions, one before/after Experiment 1. For each session, the first and second half parts were displayed in different conditions selected from XY and PTU conditions, with an overlap (40 70[sec]), i.e. 2nd half part started a little before the end of the 1st part, to give the viewers the chance of a direct comparison between XY and PTU conditions within the same scene. For each viewer, the order of display conditions and the order of conversation sessions were mixed.

Questionnaire
After viewing two sessions, subjects were requested to write down their impression, evaluations, and comparison of the different conditions, on an A4-size free-description form, within 10 minutes.

Result and Discussion
The viewers’ reports provided rich descriptions of their impressions and candid evaluations. Summarizing the viewers’ reports, the proposed system, i.e. XY condition, can be characterized as offering rich emotional expressibility, immersion in conversations, a strong sense of social presence, and potential behavioral/emotional contagion.

First, comments on the emotional expressibility in XY condition include
- “more emotionally expressive method”;
- “deep nodding looks weighty”, and
- “backward neck motion well expressed impressed gestures”.

Second, comments on the immersion in conversations and the strong sense of social presence were reported such as
- “I felt as if I was in the conversation”;
- “Despite the video, the reality and presence were fun as if the people actually there”, and
- “I felt as if people were actually sitting close to me”.

Furthermore, some subjects pointed out potential behavioral/emotional contagion in the XY condition, such as
- “I was nearly started nodding myself”;
- “I felt I was feeling empathy with persons in the conversations by watching moving panel in nodding direction”, and
- “I felt that my gaze and face direction were unconsciously moving in the direction the panel moved”.

In addition, some subjects reported that the experiment also had an impact on their perspective on understanding conversations, as indicated by
- “I realized that nodding is as importance as voice in conversations,” and
- “It was interesting that the different conditions showed me different ways of gazing”.

On the other hand, some viewers reported negative effects of the physical motion, such as
- “Overemphasized motion could give unnatural impression”;
- “Sometimes I couldn’t concentrate on the conversation because the motion disturbed me”;
- “Sometimes it was hard to distinguish the intended gesture from others like straightening one’s posture”.
- “I felt fatigued after viewing for long periods”.

The authors believe that these negative side comments have an important implication. From experiments and observations up to now, the positive effect of 4-DoF physical motion augmentation is the increased reality of reconstructed meeting scenes, in terms of emotional expressibility, social presence, immersion, and behavioral/emotional contagion. On the other hand, the negative effects may include high cognitive load for interpreting the physical motions. From the optimistic point of view, both effects are thought to be different sides of the same coin, i.e. experiencing highly realistic scenes could require the same mental efforts demanded by real meetings, especially in highly emotional and social situations.

In addition, this issue could raise the new problem of optimal motion control of the 4-DoF kinetic displays so as to well balance a number of factors, such as realistic social presence, level of emotional expressiveness, clarity of nonverbal expression such as gaze directions, and cognitive load. For example, businesslike briefings will not require highly emotional expressions, and head rotation would give sufficient information for enhancing gaze directions, for better understanding the situation of the meeting. On the other hand, for more personal meetings with friends and/or family, more emotionally rich expressions would be welcomed. In addition, there is the possibility that the conversation space itself can be adaptively controlled by using dynamic physical motion augmentation, e.g. balancing the motions among participants depending on their emotions, such as attenuating the motion of an angry participant, and so on.

3.4 Experiment 3
Experiment 3 aims to evaluate the sound reproduction of our system, which was common among the three conditions in Experiments 1 and 2. In particular, we intended to verify the feasibility of newly introduced GMM-based flat panel loudspeaker system, which is integrated into the screen panel.

Subjects
The same subjects of Experiment 1 and 2 were employed.

Questionnaire
After all experiment sessions, subjects were requested to choose one of a 7-point Likert scale to give their impression about sound, in terms of i)naturalness, ii)clarity, iii)speaker localization, iv)spatial consistency between image and voice, v)directionality changes with panel motion, and iv)awareness of mechanical noise. The experiments in this paper employed the same audio device and setting described in 2.

Result and Discussion
The average scores for i)naturalness, ii)clarity, iii)speaker localization, iv)spatial consistency between image and voice, v)directionality changes with panel motion, and iv)awareness of mechanical noise, were 2.00 (0.71), 2.00 (0.71), 1.67 (0.71), 2.22 (0.67), 1.56 (1.01), and 2.13 (0.83), respectively, where values inside parentheses are the standard deviations. The Likert score of 2.0 corresponds to “agree”. These results indicate that our GMM-based flat panel speaker is suf-
fourthly clear and has solved the problem with MM-Space, the inconsistency between facial location (on the screen) and sound source (loudspeaker on the table). On the other hand, mechanical noise remains clearly noticeable and something disturbing, as also reported by the viewers in Experiment 2. The noise is mainly generated by PTU gears, especially the nodding motion. To ease this problem, the use of direct-drive rotational motors would be advisable.

4. DISCUSSIONS AND CONCLUSIONS

This paper proposed MM+Space for recreating multiparty face-to-face conversation scenes. The main feature is the 4-DoF kinetic display, which physically augments the reproduction of human head motions. Compared to MM-Space, which has only 2-DoF rotation, MM+Space offers increased power of expressibility not only in terms of gaze direction, but also emotional interaction. Experiments showed that the 4-DoF motions significantly outperformed 2-DoF rotations in terms of the increased perception of people’s presence and the expression of their postures. In addition, it was reported that with the proposed system viewers experienced rich emotional expressibility, a stronger sense of social telepresence, immersion in the conversations, and possible behavioral/emotional contagion. Experiments implied that the 4-DoF kinetic display is characterized by not just a greater level of modality, but also a qualitative leap in rich emotional media.

Future work includes measuring the performance of the system as a communication media, and to answer questions such, what kind of and how much nonverbal information can be expressed and perceived by viewers, such as visual focus-of-attention, back-channeling, and interpersonal emotions. Furthermore, for real application, it is necessary to measure the communication quality and comprehension that the users can experience as a whole. To this end, it is important to develop proper evaluation measures that can well quantify these properties.

Moreover, authors have been building a real-time system for connecting remote places, in a series of studies on MM-Space. The major challenge is the delay problem, which occurs in sensing, transmitting, and controlling the physical motion. In particular, a novel issue is the additional delay imposed by moving physical objects, i.e. screen panels; the noticeable time gap between displayed image and physical motions left the viewers with an unnatural impression.

5. REFERENCES