Robot Motion Generation and Auditory Scene Manipulation for Tele-presence Systems

CHAORAN LIU

MARCH 2015

Robot Motion Generation and Auditory Scene Manipulation for Tele-presence Systems

A dissertation submitted to

THE GRADUATE SCHOOL OF ENGINEERING SCIENCE

OSAKA UNIVERSITY

In partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ENGINEERING

BY

CHAORAN LIU

MARCH 2015

ABSTRACT

This thesis presents a tele-operation robot system that focused on tele-presence for both robot and tele-operator. In proposed system, robot's motion was generated by a rule-based model based on dialog act function tags, auditory environment in robot side is augmented and reproduced to tele-operator for increasing tele-presence.

Head motion occurs naturally and in synchrony with speech during human dialogue communication, and may carry paralinguistic information, such as intentions, attitudes and emotions. Therefore, natural-looking head motion by a robot is important for smooth human-robot interaction. Based on rules inferred from analyses of the relationship between head motion and dialogue acts, this paper proposes a model for generating head tilting and nodding, and evaluates the model using three types of humanoid robot (a very human-like android, "Geminoid F", a typical humanoid robot with less facial degrees of freedom, "Robovie R2", and a robot with a 3-axis rotatable neck and movable lips, "Telenoid R2"). Analysis of subjective scores shows that the proposed model including head tilting and nodding can generate head motion with increased naturalness compared to nodding only or directly mapping people's original motions without gaze information. We also find that an upwards motion of a robot's face can be used by robots which do not have a mouth in order to provide the appearance that utterance is taking place. Finally, we conduct an experiment in which participants act as visitors to an information desk attended by robots. As a consequence, we verify that our

generation model performs equally to directly mapping people's original motions with gaze information in terms of perceived naturalness.

In a tele-operated robot system, the reproduction of auditory scenes, conveying 3D spatial information of sound sources in the remote robot environment, is important for the transmission of remote presence to the tele-operator. We proposed a tele-presence system which is able to reproduce and manipulate the auditory scenes of a remote robot environment, based on the spatial information of human voices around the robot, matched with the operator's head orientation. In the robot side, voice sources are localized and separated by using multiple microphone arrays and human tracking technologies, while in the operator side, the operator's head movement is tracked and used to relocate the spatial positions of the separated sources. Interaction experiments with humans in the robot environment indicated that the proposed system had significantly higher accuracy rates for perceived direction of sounds, and higher subjective scores for sense of presence and listenability, compared to a baseline system using stereo binaural sounds obtained by two microphones located at the humanoid robot's ears. The effect of variable ITD (interaural time differences) in perceived distance expression of near field sound source has been investigated. Result of localization experiment showed that compare to HRTFs measured from a dummy head, calculated impulse response which implied ITD's change in near field can provide higher sense of distance with same accuracy in perceiving direction of sound in horizontal plane. We also proposed three different user interfaces for augmented auditory scene control. Evaluation results indicated higher subjective scores for sense of presence and usability in two of the interfaces (control of voice amplitudes based on virtual robot positioning, and amplification of voices in the frontal direction).

CONTENTS

LIST OF FIGURES

LIST OF NOTATION

SVD	singular value decomposition method
MUSIC	multiple signal classification
DS beamforming	delay and sum beamforming
HRTF	head related transfer function
HRIR	head related impulse response
ms	millisecond
DFT	discrete Fourier transformation
FFT	fast Fourier transformation
KEMAR	knowles. Electronics Mannequin for Acoustics Research
$\otimes$	convolution operation
$E[\boldsymbol{A}]$	expected value of **A**
$tr[\boldsymbol{A}]$	trace of **A**
$\boldsymbol{A}'$	the transpose of **A**
$\boldsymbol{A}^*$	the complex conjugate of **A**

$\boldsymbol{A}^H$	the conjugate transpose of **A**
$\boldsymbol{A}^{-1}$	the inverse of **A**
$\|\mathbf{A}\|$	norm of **A**
$\|\|\boldsymbol{A}\|\|$	frobenius norm of **A**

1. INTRODUCTION

1.1. HEAD MOTION OF HUMANOID ROBOT

To allow smooth dialogue communication between humans and robots, besides the verbal (linguistic) information, nonverbal information such as head motion is also important for expressing paralinguistic information such as intentions, attitudes and emotions, and for increasing the robot's perceived lifelikeness.

Head motion naturally occurs during speech utterances, and can be either intentional or unconscious. In the former case, head motion may carry clear meanings in communication. For example, nods are frequently used for expressing agreement, while headshakes are used for expressing disagreement. Most of the time, however, head motion is unconsciously produced. Regardless of this difference, both types of motions are somehow synchronized with the speech utterances and transmit nonverbal information.

One of our motivations for the present work is to obtain a method for generating head motion from the speech signal in order to automatically control the head motion of a tele-operated humanoid robot (such as an android). Thus, the lifelikeness of a robot could be increased by imitating a human's natural head motion, and smooth human-robot communication can be expected.

We analyzed several sessions of free dialogue conversation between Japanese speakers, and found a strong relationship between head motion and dialogue acts (including turn taking functions). Nods occurred frequently during dialogue speech, not only for expressing dialogue acts such as agreement and affirmation, but also as indicative of syntactic or semantic units, appearing at the last syllable of the phrases, with strong phrase boundaries. At weak phrase boundaries where the speaker is thinking or indicates that he/she did not finish his/her speech, head tilts were frequently observed. Also, a rule-based nod generation model was proposed and evaluated.

Furthermore, we extend our study of head motion generation by adding a head tilting generation model, based on human head motion analysis results. We then evaluate the proposed model in two types of humanoid robots: a female android, “Geminoid F”, and a typical humanoid robot with less degrees of freedom, “Robovie R2”.

The effects of an additional “face up” motion during utterances are also evaluated, with the goal of reducing perceived unnaturalness in robots that do not have a mouth (i.e., movable lips).

1.2. TELE-PRESENCE SYSTEMS

Tele-communication systems allow people communicating with others over a distance. This capability brings great potential, but also be accompanied with deficiency. One area where conventional tele-communication systems failed to deliver is presence. This absent of physical presence may impede fluency or mislead the communication. Various tele-operated communication robots are developed to fill the gap between face-to-face communication and video-audio based tele-communication since robots' body may be accounted as a carrier of physical presence.

In order to realize tele-presence through robots, most works so far have focused on evaluating the sense of presence in the robot side, i.e, how people around the robot would feel the presence of the tele-operator through the robot [1][2]. However, the sense of presence from the tele-operator viewpoint has not been given much attention in tele-presence robot systems, so far.

Considering a scenario of a robot involved in multi-party communication where the tele-operator communicates with other people through a tele-operated robot, there are deficiency of many information such as intelligibility, comfort and physical presence, in comparison to face-to-face communication. Especially the spatial information of sounds is lost when using a conventional (monaural) microphone-headphone system.

The goal of this part is to develop a tele-operation robot system which can reproduce and manipulate auditory scenes (spatial auditory sensation of sound sources) of a remote robot environment, in order to transmit sense of presence to the tele-operator. We refer such a system as "tele-presence system" hereinafter. A tele-presence system requires the ability to localize sound sources, separate them and render the sound sources at appropriate spatial locations, in real time.

In this work we proposed a tele-presence system which is able to synthesize and manipulate the auditory scenes of a remote robot environment, matched with the operator's head orientation. For that purpose, the spatial information of sound sources are captured by using multiple microphone arrays and laser range finders in the robot environment, and synthesized with HRTFs properly selected according to the sound locations in the robot environment, with the coordinates rotated according the operator's

head orientation. In addition, we designed three user interfaces for allowing augmented auditory scene control, and evaluated them through subjective experiments.

2. RELATED LITERATURE

Head motion analyses can focus on either of two problems: how to recognize a user's head motion and interpret its role in communication [3][4]; or, how to generate a robot's head motions in synchrony with the robot's speech. The present work focuses on the latter problem of generating natural head motion while the robot is speaking.

Many studies have tried to find a correspondence between head motions and prosodic features, such as the fundamental frequency (F0) contours (which represents pitch movements), and energy contours, in several languages [5-10]. For example, emphasis of a word often goes along with head nodding, and a rise of the head can correspond with a rise in voice, in English [8]. For Swedish, words with focal accent are accompanied by a greater variation of the facial parameters (including head motions) than words in non-focal positions, in all expressive modes [9]. However, it is reported in [5] that, correlation among F0 and the 6DOF (rotation and translation) of the head motions was between 0.39 and 0.52 for English speakers, and between 0.22 and 0.30 for Japanese speakers. This shows that the use of F0 alone is not enough to generate head motions, and that the correspondence between F0 and head motions is language dependent. Also, most of these studies analyse read speech or acted emotional speech data. In [11], relations

between head movements and the semantics of utterances are analysed in Japanese spoken dialogue, by also considering speaking turn and speech functions. Regarding head motion generation, a system that uses corpus-based selection strategies to specify the head and eyebrow motion of an animated talking head is presented in [12]. The system considers syntactic, prosodic and pragmatic context for generating the motions. However, only data for one speaker was analysed.

On the other hand, virtual reality has gained great popularity as a way to simulate physical presence in medical, game and military applications [13][14]. Virtual reality has also been applied for tele-conferencing/tele-operating systems [15][16]. One goal in the development of a tele-operating system is creating tele-sensation which refers to a sense of presence in a remote location. Most of virtual reality-related studies have focused on creating virtual visual sensation using immersive projection displays [16][17] or virtual reality unit (a helmet-like visual device, usually including stereo vision) [15]. On the other hand, only a few focus on manipulating spatial auditory sensation [18][19]. Virtual auditory scene reproduction has been widely applied on game applications. However, there are few or none applications on social media, such as in tele-presence robots. In virtual reality applications that aim for improving quality and ease of interaction, rich auditory sense is a factor that cannot be ignored.

The most common way to reproduce an auditory scene is playing binaural recorded sounds using stereo headphones. However, this method requires very specific recording setup, and also loses dynamic cues and the ability to rearrange sounds from different locations. To store characteristic information corresponding to spatial properties, an efficient way is using microphone array techniques. There are many teleconference-related studies using microphone arrays for localizing sound sources [20], separating the sound from a target direction [21], or suppressing echo [22]. However, the output of such systems is monaural so that auditory scene is not reproduced.

Regarding reproduction of 3D audio, by using more channels than a stereo sound, surround sound systems have been developed for providing a better spatial sensation. Directional Audio Coding (DirAC) is used in several studies for reproducing spatial sounds on multiple loudspeaker systems [18][19]. However, the use of loudspeaker systems for auditory scene reproduction raises two problems. The first one is that spatial sounds will be played in an unknown environment, thus, different configurations of room and loudspeakers will impose unknown nonlinear transformations that usually cannot be compensated. The second problem is that all loudspeaker systems have a constraint that the listener has to be in the "sweet spot" (an optimal listening position) which is limited to the central point in the system setup [23].

Another strategy is to synthesize binaural sounds using stereo headphones. It is generally accepted that human perceives the spatial position of a sound source based on the differences between the sounds at the two ears. These auditory differences include sound pressure level and arrival times at the two ears [24][25]. The binaural sounds to be reproduced with headphones can be synthesized by using measured Head Related Transfer Functions (HRTF), which are functions representing the relationship between the sound source position and the sound wave property at the two eardrums of a particular person [25].

However, two problems arise when using HRTF to represent an auditory scene through headphones in tele-operation applications. One is that when the operator moves his body or head, the auditory scene essentially rotates with the operator's movement. This effect may cause a decrease in the perceived sense of presence. In order to keep a natural auditory sensation, the virtual environment should stay fixed with a real world coordinate system instead of moving with the coordinate system inside the operator's head. In other words, the sound sources should be perceived as staying at the same places irrespective to the operator's movement.

The other problem of HRTF is front-back confusion, in which a frontal sound source is wrongly perceived as coming from behind or vice versa. In daily life, when people localize a sound, sometimes they move their head consciously or unconsciously to change the relative position of the sound. It is reported that people tend to move their head to help themselves for sound localization [26]. These movements effectively reduce the front-back error [27].

3. ROBOT HEAD MOTION GENERATION

3.1. MOTION DATA COLLECTION

The motion capture system used is the Hawk system from Motion Analysis. Ten infra-red cameras are arranged in a rectangle around a room in order to capture the motions of both speakers. Seven hemispherical passive reflective markers are applied to the speaker's head, nose and chin, as shown in Figure 3.1. The markers on the head and nose provide a static reference frame for the (rigid) head, while the marker on the chin (relative to the nose marker) was used to align the motion data with the speech audio data, since systematic errors sometimes occurred in the synchronization between these data.

It is worth mentioning that although the situation with motion capture markers would be experienced as unnatural, all speakers agreed that this unnatural feeling happened only in the beginning of the first time they experienced the motion capture markers. After a while (say, 1 or 2 minutes), they simply forget about the markers, and can talk naturally.

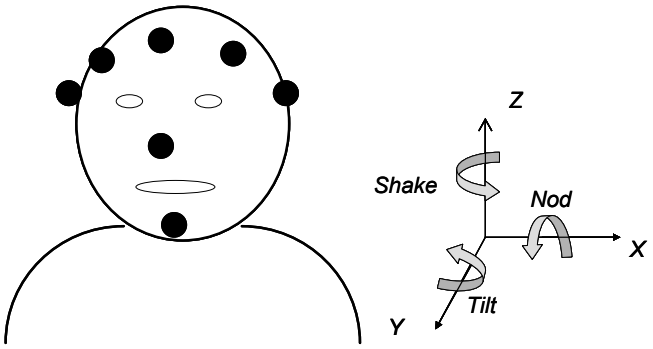


Figure 3.1. Markers and angles used to describe head motions.

The three rotation angles, shown in Figure 3.1, are used to describe the head motions. We use the terms "nod", "shake" and "tilt", in correspondence with the terms "pitch", "yaw", and "roll", used in aerodynamics. The head rotation angles are estimated from the markers based on the singular value decomposition method (SVD) [28].

We used six motion capture markers (excluding chin marker) for calculating head rotation. The initial position of this marker set is represented as:

$$\boldsymbol{P}_{ini} = [x_1, x_2, \dots x_6; y_1, y_2, \dots y_6; z_1, z_2, \dots z_6]^T$$

and the current position is:

$$\boldsymbol{P}_{cur} = [x'_1, x'_2, \dots x'_6; y'_1, y'_2, \dots y'_6; z'_1, z'_2, \dots z'_6]^T$$

then the rotation matrix **R** is such that:

$$\boldsymbol{P}_{cur} = \boldsymbol{R}\boldsymbol{P}_{ini}$$

can be computed as:

$$\boldsymbol{R} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^T$$

where **U** and **V** are such that:

$$[\boldsymbol{U}, \boldsymbol{D}, \boldsymbol{V}^T] = svd(\boldsymbol{P}_{ini}^T \boldsymbol{P}_{cur})$$

$\boldsymbol{D}$ is diagonal matrix include singular values. $\boldsymbol{S}$ is also diagonal matrix with $\boldsymbol{S}_{ii}=\pm 1$. The signs are taken to be the same as the corresponding diagonal elements of the $\boldsymbol{D}$ matrix.

Initial position $\boldsymbol{P}_{ini}$ and current position $\boldsymbol{P}_{cur}$ and target are the 3D marker set of the neutral and current positions respectively, translated to new coordinates having their centroids as origin. The neutral positions are obtained in intervals where subjects are looking directly ahead. The rotation matrix is obtained by:

The rotation angles are then obtained from the elements of the rotation matrix $\boldsymbol{R}$ according to the following expressions:

$$Angle_{tilt} = \tan^{-1}\left(\frac{\boldsymbol{R}_{(2,1)}}{\boldsymbol{R}_{(1,1)}}\right)$$

$$Angle_{nod} = \tan^{-1}\left(\frac{-\boldsymbol{R}_{(3,1)}}{\sqrt{\boldsymbol{R}_{(3,2)}{}^{2} + \boldsymbol{R}_{(3,3)}{}^{2}}}\right)$$

$$Angle_{shake} = \tan^{-1}\left(\frac{\boldsymbol{R}_{(3,2)}}{\boldsymbol{R}_{(3,3)}}\right)$$

Collected head motion data were used for relationship analysing and reproducing human movement on humanoid robot during evaluation experiments.

3.2. DIALOG ACT

In [29], strong relationships between head motion and dialogue acts (including turn taking functions), and between dialogue acts and prosodic features, were reported. In the present work, we focus on the relationship between dialogue acts and head motion.

Dialogue act tags have been annotated for each phrase in a database of dialogue between several pair of speakers, according to the following set, based on the tags proposed in [30], taking into account dialogue acts such as affirmative or negative reaction, expression of emotions like surprise or unexpectedness, and turn-taking functions.

- k (keep): the speaker is keeping the turn; a short pause or a clear pitch reset is accompanied at strong phrase boundaries.
- k2 (keep): weak phrase boundaries in the middle of an utterance (when no pause exists between phrases).
- k3 (keep): the speaker lengthens the end of the phrase, usually when thinking, but keeping the turn (may or may not be followed by a pause).
- f (filler): the speaker is thinking or preparing the next utterance, e.g., “uuun”, “eeee”, “eettoo”, “anoo” (“uhmmm”).
- f2 (conjunctions): can be considered as non-lengthened fillers, e.g., “dakara”, “jaa”, “dee” (“I mean”, “so”).
- g (give): the speaker finished talking and is giving the turn to the interlocutor.
- q (question): the speaker is making a question or asking for a confirmation to the interlocutor.

- bc (backchannels): the speaker is producing backchannels (agreeable responses) to the interlocutor, e.g., “un”, “hai” (“uh-huh”, “yes”).
- su (admiration/surprise/unexpectedness): the speaker is producing an expressive reaction (admiration, surprise) to the interlocutor’s utterances, e.g., “heee”, “uso!”, “ah!” (“wow”, “really?”)
- dn (denial, negation): E.g., “iie”, and “uun” accompanied by a fall-rise pitch movement (“no”, “uh-uh”).

Among the analysis results for the relationship between head motion and dialogue acts based on dialogue between several pairs of speakers, it was shown that nodding is the head motion that most frequently occurs during dialogue speech, and that it frequently appears in backchannels (bc) and at the last syllable of strong phrase boundaries (k, g, q). At the weak phrase boundaries where the speaker is thinking, embarrassed, or indicates his/her speech utterance has not been concluded (f, k3), head tilting was frequently observed.

In the present work, we exploit these analysis results for creating head motion generation models.

3.3. HEAD MOTION GENERATION

We start proposing a very simple nodding model of controlling only the timing of the nods. In this model, nods are generated in the centre of the last syllable of utterances with strong phrase boundaries (k, g, q) and backchannels (bc). The utterance segmentation and the respective dialogue act tags were used for providing the timing of the nods.

Furthermore, our rule-based head motion generation model for nodding was extended to include head tilting based on analysis results, and the effects of head tilting control in two types of humanoid robot were evaluated.

In this proposed model, nods are generated in the center of the last syllable of utterances with strong phrase boundaries (k, g, q) and backchannels (bc), while head tilts are generated in the weak phrase boundaries where the speaker lengthens the end of the phrase or put a pause in the middle of a sentence due to disfluencies (f, k3). The utterance segmentation and the respective dialogue act tags were used for providing the timing of the nods.

Figure 3.2 shows examples of several shapes of nod and head tilt found in the database. The duration of the nods varied around 0.4 to 0.7 seconds. Note the presence of a slight upward motion, which often occurs before the characteristic down-up motion of nods. The duration of the head tilt samples varied in time from 0.8 to 1.5s, and was more dependent on the phrase length.

Figure 3.3 shows representative nod and head tilt shapes, extracted from the database, which are used in the motion generation.

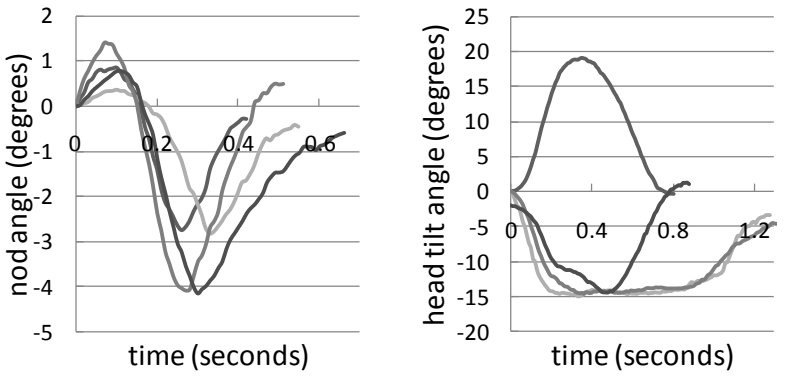


Figure 3.2. Examples of observed nod and head tilt shapes, extracted from the database.

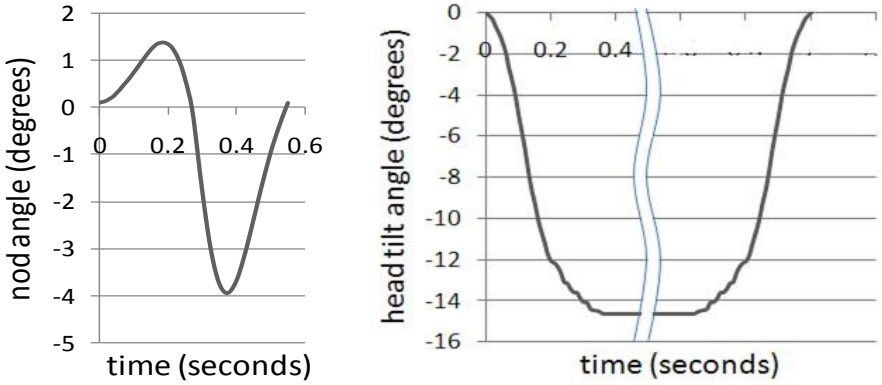


Figure 3.3. Representative nod and head tilt shapes used in the head motion generation model.

Fixed shapes for single nods and head tilts were used (i.e., the intensity and the duration of the nods, and the intensity of head tilts were kept the same), so that the effects of timing can be evaluated. Regarding head tilt, the tilt angle (of 15 degrees) was kept until the phrase finished, so that the length of the motion was decided according to the inter-phrase interval lengths.

3.4. EVALUATION OF NOD AND HEAD TILT

3.4.1. EXPERIMENTAL SETUP

Eleven conversation passages with durations between 10 to 20 seconds, including fillers and turn keeping functions (f and k3), were randomly selected from our database, and rotation angles (nod, shake and tilt angles) were extracted for each utterance. The duration of the conversation passages was limited to 10 to 20 seconds, since subjects have to compare a pair of motions for the same speech utterances, and such comparison would be difficult if each stimulus is too long. Also, as a dialogue act is attributed for each phrase unit, utterances of 10 to 20 seconds usually contain more than 10 phrases, i.e., some context information is still present in the dialogue passage.

Head rotation angles were computed by the head motion generation model described in the previous sub-section (NOD&TILT) for each conversation passage.

For comparison, we prepared two types of motion. One is a reproduction of the head rotations extracted from the original motion capture data (ORIGINAL). Another is the nod only motion (NOD ONLY).

The motions were generated in two types of robots; one is a female android robot (Geminoid F) and the other is a humanoid robot (Robovie R2), as shown in Figure 3.4.

Figure 3.4. External appearance of Geminoid F and Robovie R2.

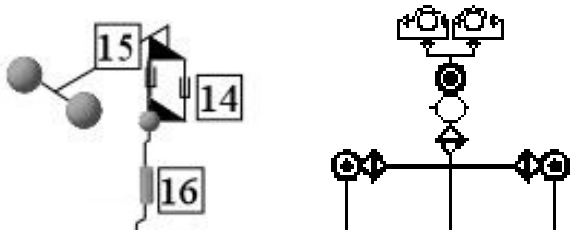


Figure 3.5. Head actuators for Geminoid F and Robovie R2.

Robovie R2 has three degrees of freedom for its head, so that the rotation angles could be directly mapped to the actuator commands by a linear mapping. Geminoid F also has three actuators for the head (as shown in Figure 3.5), but their correspondence to the three rotation angles is not as straight-forward as for Robovie R2. A mapping function between the rotation angles and the android actuator commands proposed in previous work was therefore applied [29]. Actuators 14 and 15 (left panel in Figure 3.5) move the head, when they are controlled independently, in a slantwise direction from left down to right up, and right down to left up directions, respectively. Nod (vertical) motion can be conducted by moving actuators 14 and 15 simultaneously with the same command value. Actuators 16 (left panel in Figure 3.5) moves the head around yaw axis from left to right. In order to compute head angles, a hand-selected human head position is required in addition as neutral position. We needed this neutral head position as references to calculate the angle for every frame during the speech. The actuation values for each actuator are given by:

$$act_{14} = act_{neutral} + \frac{angle_x}{max\_angle_x} \cdot 127 + \frac{angle_y}{max\_angle_y} \cdot 127$$

$$\mathrm{act_{15}} = \mathrm{act_{neutral}} + \frac{\mathrm{angle_x}}{\mathrm{max\_angle_x}} \cdot 127 - \frac{\mathrm{angle_y}}{\mathrm{max\_angle_y}} \cdot 127$$

$$\mathrm{act_{16}} = \mathrm{act_{neutral}} + \frac{\mathrm{angle_z}}{\mathrm{max\_angle_z}} \cdot 78$$

where $act_{neutral}$ is the activation value for the neutral position (127 for all actuators), *angle* is the target rotation angle, and *max_angle* is the maximum rotation angle of android robot.

The actuation values obtained above are clipped to the limit ranges for each actuator.

Commands are sent each 20 ms for Geminoid F, and each 100 ms for Robovie R2, due to constraints in the current robot's hardware, but these rates are thought to be enough for head motion control purposes. Furthermore, for the android, the lip motion (jaw lowering motion) was reproduced from the distances between the nose and chin markers in the original motions.

Video clips were recorded for each stimulus, resulting in 33 stimuli (11 conversation passages and 3 motion types) for each robot type. For each trial, subjects saw a sequence of two videos with different head motion control types for the same conversation passage, so that they could compare the effects of changing the head motion control.

Pairs of stimuli were presented to subjects in the following order:

- **NOD ONLY vs. NOD&TILT**
- **NOD&TILT vs. ORIGINAL**
- **NOD ONLY vs. ORIGINAL**

The video pairs were sorted in random order.

Subjects were asked to rate the naturalness of the motion for each stimuli, and preference scores for each pair of stimuli, according to the following questionnaire. Only one item was allowed to be chosen for each question.

- **Is the motion of the robot natural?**

  clearly unnatural (1)

unnatural (2)

slightly unnatural (3)

difficult to decide (4)

slightly natural (5)

natural (6)

clearly natural (7)

- **Comparing the two stimuli, which one is more natural?**

the first is clearly more natural

the first is more natural

the first is slightly more natural

difficult to decide (both are natural or unnatural)

the second is slightly more natural

the second is more natural

the second is clearly more natural

38 paid subjects (18 male and 20 female) participated in the experiment. All subjects were not involved in robotics research, and ages ranged from 18 to 60 (average=34, stdev=13).

3.4.2. EXPERIMENTAL RESULTS

Figure 3.6 shows average subjective naturalness scores for individual motion types, and for each robot type. Subjective naturalness was quantified using a scale of 1 to 7.

To understand the significance of these scores, a repeated measures analysis of variance (ANOVA) was conducted. A significant main effect was found ($F(2,36) = 26.152$, $p < 0.0005$ for Geminoid F and $F(2,23) = 19.109$, $p < 0.0005$ for Robovie R2).

Regarding the naturalness of individual motion, NOD&TILT was judged to be the most natural, while NOD ONLY and ORIGINAL obtained lower scores in both robot types. Possible reasons for these results are discussed in Section 3.5.

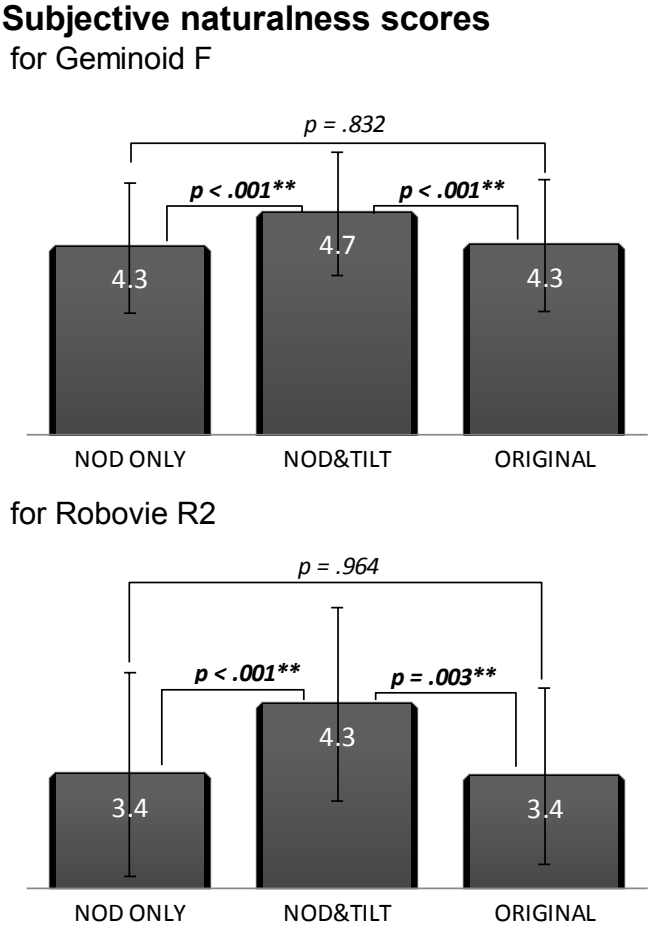


Figure 3.6. Subjective naturalness for each motion type (standardized value), for Geminoid F (left) and Robovie R2 (right). Error bars indicate Standard Deviation.

Comparing the two graphs in Figure 3.6, for all three motion types, scores for Geminoid F were higher than for Robovie R2. As quite similar motion was reproduced in the two types of robots, we believe the difference of external appearance causes this discrepancy. This is discussed in more detail in Section 3.4.3.

For Geminoid F, comparison between "NOD ONLY" and "NOD&TILT" shows that "NOD&TILT" (single nods plus head tilts with adequate timing) is significantly more natural than "NOD ONLY" ($p < 0.0005$), indicating the effectiveness of the head tilt motion generated in the specific weak phrase boundaries. Comparison between "NOD&TILT" and "ORIGINAL" showed that "NOD&TILT" model is more natural than "ORIGINAL" ($p < 0.0005$). For "NOD ONLY" and "ORIGINAL", it was expected that "ORIGINAL" would have higher preference scores. However, the results do not show a significant difference between these two cases. This means that simply trying to reproduce the original motion does not necessarily imply that natural motion will be generated.

For Robovie R2, we can observe that there were similar overall tendencies to the results for Geminoid F. "NOD&TILT" is evaluated significantly higher than "NOD ONLY" and "ORIGINAL" ("NOD&TILT vs. "NOD ONLY": $p < 0.0005$; vs. "ORIGINAL": $p = 0.003$).

We think a better control of the other rotation angles and the intensity of nodding may improve naturalness. However, the present results show that a "slightly natural" motion could be achieved even by using a very simple timing control model.

3.4.3. FURTHER ANALYSIS

Regarding the differences in the subjective naturalness scores between the two robot types, the results in Section 4.4.2 showed that when the same motion was used in both robots, the scores for Robovie R2 were lower score than for Geminoid F. One possible reason could be that Robovie R2 does not have movable lips. For the android, lip motion was reproduced from the distances between the nose and chin markers, so that the subjects can recognize the intervals of the speech utterance through both visual and auditory information. However, for Robovie R2, as head motion is not strongly related to phonetic features, it is difficult to feel that the speech is being uttered by the robot by

only watching the robot's head motion. We think that this lack of visual information causes the lower scores in subjective naturalness for Robovie R2, as it does not have movable lips.

3.5. EVALUATION OF FACE UP MOTION FOR UTTERANCE SIGN

To address the problem discussed in Section 4.3.3 of increasing perceived naturalness for robots without movable lips (such as Robovie R2), we propose a slight "face up" motion (3 degrees) during speech utterance intervals, as an indication for utterance intervals. This upward motion is also often observed in natural speech, and should improve motion naturalness in a robot which does not have movable lips, such as Robovie R2.

3.5.1. EXPERIMENTAL SETUP

The head motion generation models "NOD ONLY" and "NOD&TILT" described in previous section 3.1 were combined with a face up motion during speech utterance intervals. We call these new generation models "NOD ONLY+" and "NOD&TILT+".

8 conversation passages were randomly selected from the previous experiment. For each conversation passage, six types of motion were generated, comprising "NOD ONLY+" and "NOD&TILT+" for Robovie R2 as well as "NOD ONLY" and "NOD&TILT" for both Robovie R2 and Geminoid F (the latter were used as baseline conditions for comparison).

Video clips were recorded for each speech segment accompanied by motion, resulting in 48 stimuli for Robovie R2 and Geminoid F. As in the previous experiment (in Section 3), subjects saw a sequence of two videos with different head motion control types for the same conversation passage, and were asked to rate the naturalness of the motion for each segment.

10 Japanese subjects (5 male and 5 female) participated in the experiment. All subjects were participants in their 20s (college students) not involved with robotics research.

3.5.2. EXPERIMENTAL RESULTS

Figure 3.7 shows subjective naturalness scores for each motion type. As in the previous experiment, subjective naturalness was quantified using scale numbers of 1 to 7 where "1" was described as "clearly unnatural" and "7" was described "completely natural".

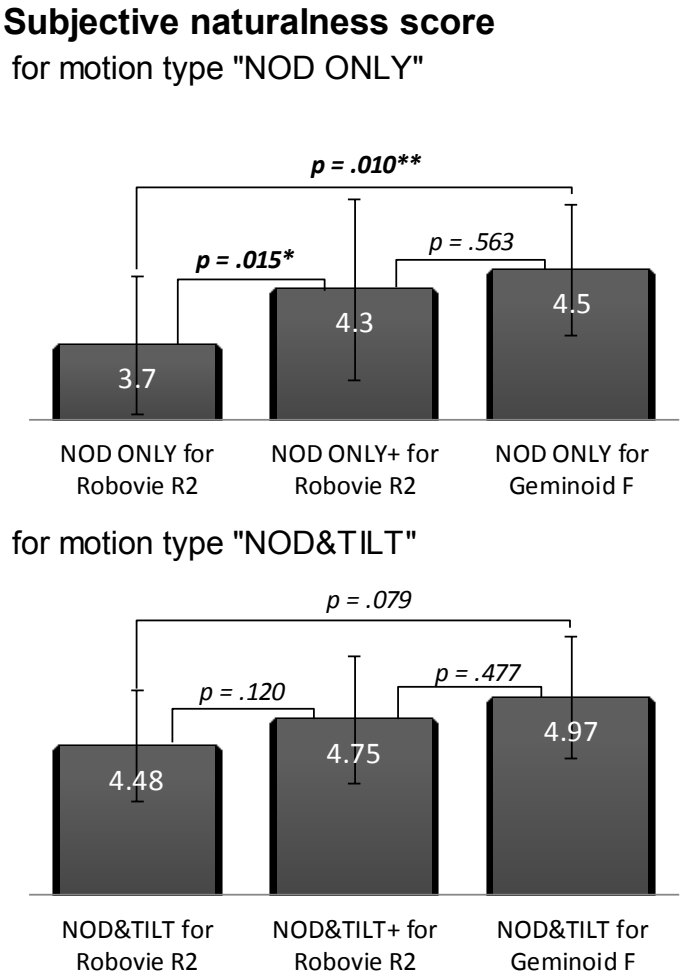


Figure 3.7. Subjective naturalness for NOD&TILT+, compared with NOD ONLY model and NOD&TILT model. Error bars indicate Standard Deviation.

We compared the scores for motion types with upward motion (Robovie R2) against those without upward motion in both robots (the 2 baseline conditions). The middle bar in Figure 3.7 shows the subjective naturalness scores for the proposed "NOD&TILT+" model (with face up motion) for Robovie R2, while the left and right bars show the naturalness scores for "NOD&TILT" (without the face up motion) model for Robovie R2 and Geminoid F, respectively.

In place of lip movements synchronized to speech utterance, the face up motion during the speech utterance is expected to have a similar effect.

Analysis revealed a significant difference between "NOD ONLY+" and "NOD ONLY" for Robovie R2. However, a significant difference was not found between "NOD&TILT+" and "NOD&TILT" for the same robot. The reason for this is that the head tilt during the speech utterance served much the same purpose as the face up motion by attracting the listener's attention before the final nod. It could be interesting in the future to further clarify the cause for these results and to investigate if there are other motions which can also be used in this way to improve perceived naturalness.

However, the face up motion was still not enough to make the subjective naturalness scores for Robovie R2 as high as for Geminoid F. As Geminoid F is a very human-like anthropomorphic robot, which is at first glance indistinguishable from real humans, external appearance could make the subjective naturalness results better. We plan to evaluate the effect of face up motion on other types of robots in the future.

3.6. EVALUATION OF HEAD MOTION AND GAZE DURING HUMAN-ROBOT COMMUNICATION

The results from the experiment in Section 3.4.2 left questions unanswered. First, how do subjects evaluate the models when they are interacting with a real robot and not merely watching video footage? And, why was the "ORIGINAL" motion, which we thought would be perceived as most natural, judged to be as unnatural as the "NOD ONLY" model and less natural than the proposed "NOD&TILT" model? For the latter question, the literature strongly points to the importance of gaze movements for a wide variety of interaction schemes, both in human science [31][32] and for HRI [33][34]. However, in the "ORIGINAL" motion, only head movements, with no gaze movement, were reproduced.

Therefore, we conducted an experiment with two purposes: to verify the goodness of our models when actual robots are used and to see if the addition of gaze movements would give the expected results for original motion.

3.6.1. ANALYSIS OF GAZE

We first analyzed eye gazing in human-human dialogue conversations. Thirteen sessions of free conversations in our multimodal dialogue speech database were analyzed. The following tag set was used to annotate eye gazing based on the displayed video information from the speaker's viewpoint.

- u: upwards gaze;
- d: downwards gaze;
- l: to the left (from speaker's viewpoint);
- r: to the right;
- ul: up and to the left;

- ur: up and to the right;
- dl: down and to the left;
- dr: down and to the right;
- no: no gaze movement; (looking at conversation partner)

Figure 3.8 shows the distributions of the eye gaze tags for each dialogue act tag (described in Section 2), i.e., where conversation partners looked during the dialogue when speaking. The intensity of each square indicates how often speakers looked in that direction. The darker the background-color of the square the higher the frequency of that eye gaze act.

In Figure 3.8, it can be observed that in most cases (especially for “bc”, “g”, and “k”), the person speaking mostly looked directly toward the listener. However, the speaker looked away in about 50% of the cases when lengthening their previous utterance and thinking about what to say next (“f” and “k3”). In such cases, the speaker had a tendency to look downwards rather than upwards, when looking away. For “k3”, speakers also looked frequently down and to the right (19% of the time).

The results in Figure 3.8 indicate that eye gazing has similarities with head tilting. In both cases, behavior reflects weak phrase boundaries when a speaker lengthens the end of the phrase or put a pause in the middle of a sentence due to disfluencies (f, k3). Figure 3.9 also shows a physical relationship (interdependency) between head tilting and eye gazing.

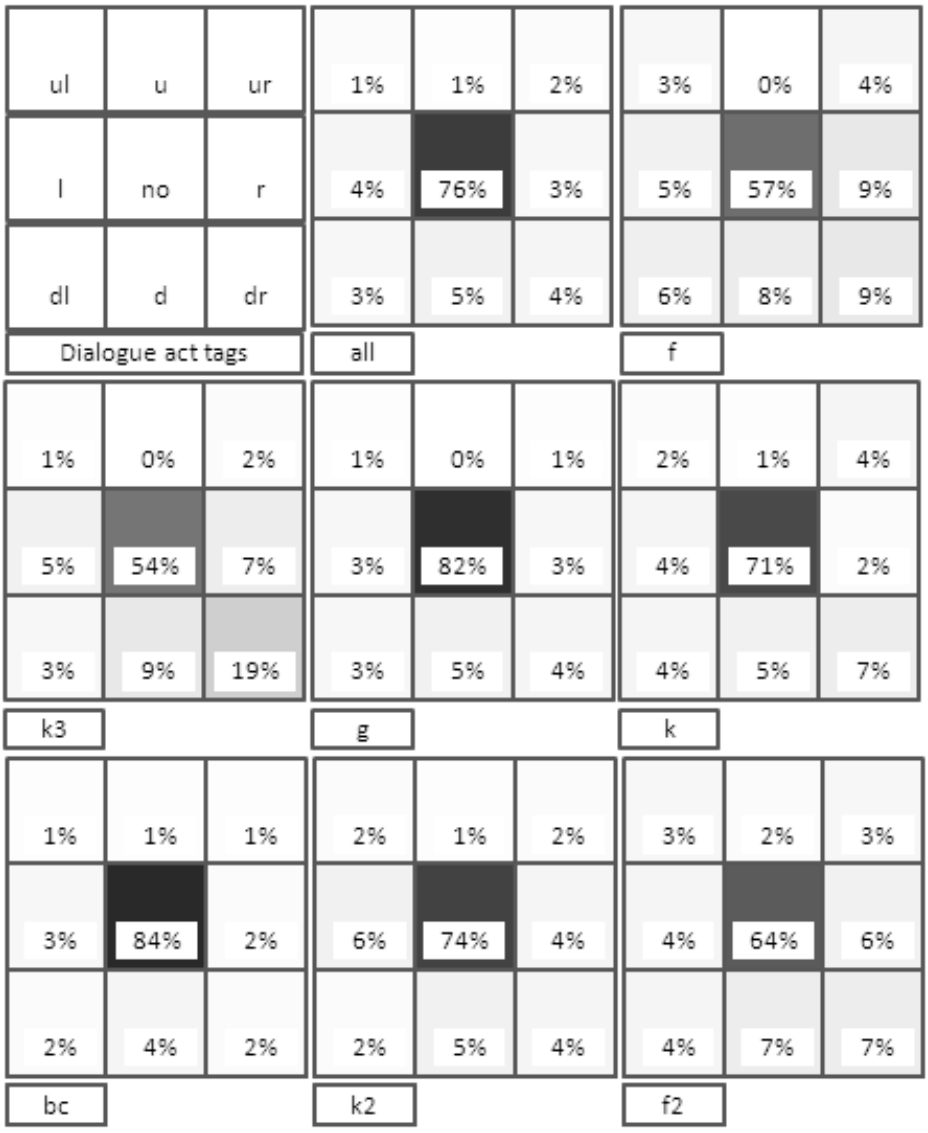


Figure 3.8. Distributions of the eye gaze tags for different dialogue acts: intensity for each square indicates how often speakers looked in that direction.

During head tilting, a speaker most often looked directly at the listener (56%), but also often looked in the direction opposite to the direction in which their head was tilted (25%). (E.g., if the speaker tilted their head left, they often looked to the right.)

Sometimes the speaker also looked in the same direction as the head tilt (10%), or downwards (8%). The speaker rarely looked upward (1%).

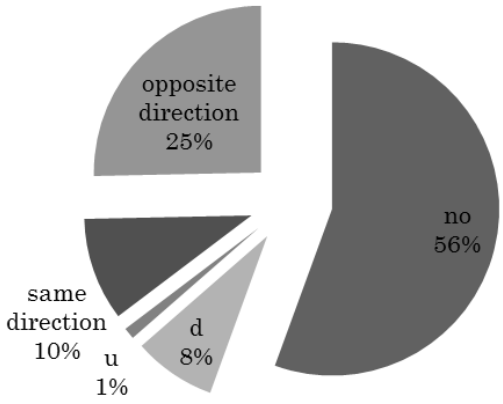


Figure 3.9. Gaze direction during speaker head tilting: "same direction" ("opposite direction") means gaze direction coincides (does not coincide) with the direction of the head tilt.

The above results indicate that if a robot alters its gaze direction during a head tilt by looking in the opposite direction, this could be perceived as more natural. Based on this, we decided to also evaluate the control of gaze in addition to the head motion in the human-robot interaction experiment of the present section.

3.6.2. EXPERIMENTAL SETUP

In order to allow for comparison of different motion generation strategies in face-to-face human-robot interaction, we created a scenario in which subjects act as visitors to an information desk and robots play the role of receptionists. In this scenario, we can ask subjects to make the same questions to the robots, so that the effects of different motion generation strategies can be fairly compared, by using the same utterances with same lip motion and changing only the head motion and gazing strategies.

We first simulated a human-human interaction for the scenario above, and recorded audio and motion capture data for a female speaker playing the role of

receptionist (collaborator). During the interaction, another speaker playing the role of the visitor asked several questions in order to induce the receptionist to naturally produce thinking behavior. The following items appeared in the interaction:

1. directions to IRC Lab.
2. what kind of research is done at IRC Lab.
3. how many staff work at IRC lab.
4. directions to the bathroom.
5. places to eat
6. Japanese restaurants
7. timetable for the bus
8. directions to Kobe
9. sightseeing spots in the vicinity
10. sightseeing spots in Nara

Two robots were used for the experiment: an android robot, Geminoid F, and a child-sized minimally-designed robot Telenoid R2 (see Figure 3.10). Telenoid R2, like Robovie R2, has a head rotatable about three independent axes and is less humanlike in appearance than Geminoid F and, but also has movable lips, which allows for clearer comparison of motion types.

Figure 3.10. External appearance of Telenoid R2

At the start of the experiment, subjects were given instructions and told they would be playing the role of a visitor to an information desk where the two robots played the role of receptionists. The subjects then entered a room in which the two robots were positioned around a desk, and met and conversed with each robot one at a time. Specifically, subjects asked each robot the same questions in the list above, in a natural manner. After each question, the robot responded according to a remote control by the experimenters. This was conducted five times, for each of the five motion types described below. After each session, subjects were asked to fill out a brief questionnaire. Answers for possible questions were prepared ahead of time by asking a collaborator to act out the scenario; audio, video and motion data of our collaborator were recorded and used as the original data to create the robot's answers.

Five motion control types were reproduced on the two robots.

- **NOD ONLY**
- **NOD&TILT**
- **NOD&TILT with GAZING**
- **ORIGINAL**
- **ORIGINAL with GAZING**

Motion control types "NOD ONLY", "NOD&TILT", and "ORIGINAL" are the same as described in Section 3 (the latter is generated by reproducing original motion from motion capture data).

During the analysis of eye gaze movements in Section 5.1, we found that when people tilt their heads, they often turn their gazes in the opposite direction. We applied this rule to generate eye gaze movements for the model 'NOD&TILT', thereby creating a new condition, "NOD&TILT with GAZING".

Eye gaze control for "ORIGINAL with GAZING" was based on annotated eye gazing tags corresponding to each moment.

For all five motion control types, lip motion was generated using a method proposed in [35], which is based on a rotation of the vowel space given by the first and second formants around the center vowel, and a mapping to the lip opening degrees.

Motion types were presented in a nearly random fashion, as we wished to have all subjects compare 3 types directly: "NOD&TILT", "NOD&TILT with GAZING", and "ORIGINAL with GAZING". Therefore, these 3 motion types were always presented together as a group either in this order, or reversed. The order of this group and the other two motion types was then determined randomly.

Subjects were asked to fill out a questionnaire for each motion control type to evaluate the naturalness of the robot's motion. The questionnaire had the following items:

- **Is the motion of the robot natural?**
- **Comparing with the previous one, which one is more natural?**

The options for the answers to these questions were the same as described in Section 3.2.

A total of 22 paid Japanese speakers (11 male and 11 female, aging from 18 to 50, average=29, stdev=11) participated in the experiment.

3.6.3. EXPERIMENTAL RESULTS

Figure 3.11 shows subjective naturalness for individual motion types, and for each robot type. The results of subjective naturalness were quantified by using a 1 to 7 scale where "1" was described as "clearly unnatural" and "7" was described as "completely natural".

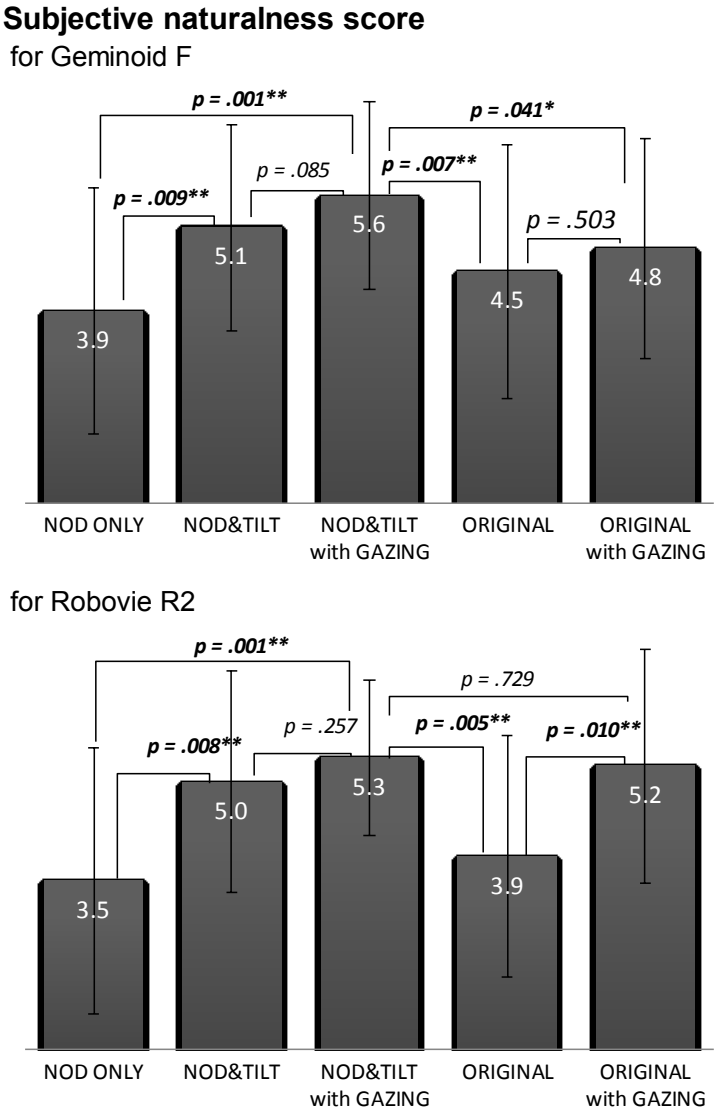


Figure 3.11. Subjective naturalness for each motion type and robot: Geminoid F (top) and Telenoid R2 (bottom). Error bars indicate Standard Deviation.

Comparing the performance of the motion control type using repeated measures ANOVA gave $F(4,18) = 4.451$, $p < 0.011$ for Geminoid F, and $F(4,15) = 4.558$, $p < 0.013$ for Telenoid R2.

We can see that similar results were observed for motion control types “NOD ONLY”, “NOD&TILT” and “ORIGINAL” as were observed previously in the video-based evaluation in Section 3.4.

Both for the Geminoid F and Telenoid R2, subjective scores for motion type “NOD&TILT with GAZING” are significantly higher than scores for motion types “NOD ONLY” and “ORIGINAL” (“NOD&TILT with GAZING” v.s. “NOD ONLY”: $p = 0.001$ for Geminoid F, $p = 0.001$ for Telenoid R2; “NOD&TILT with GAZING” v.s. “ORIGINAL”: $p = 0.007$ for Geminoid F, $p = 0.005$ for Telenoid), but comparison with “NOD&TILT” only shows non-significant or almost-significant differences ($p = 0.085$ for Geminoid F, $p = 0.257$ for Telenoid R2). This result is discussed in the next section..

For the Telenoid, “ORIGINAL with GAZING” was evaluated as highly as “NOD&TILT with GAZING” and significantly higher than “ORIGINAL” ($p = 0.010$) as expected. However, for the android, this was not the case; “NOD&TILT with GAZING” performed best. A comparisons between “ORIGINAL with GAZING” and “NOD&TILT with GAZING” shows significant difference ($p = 0.041$). We suspected that imperfect reproduction of the original motions was again the culprit. Specifically, Geminoid F cannot tilt her head about the roll axis without also rotating in the pitch axis due to the position of her head actuators (see Figure 3.5), causing her chin to move forwards. To confirm this, we recorded the robot’s reproduced motion data using the motion capture system, and compared it with the original motion data which were captured from a human speaker. We found that pitch and yaw motions were reproduced well, but that roll motions were not accurately reproduced, causing a potential source for perceived unnaturalness (see Figure 3.12).

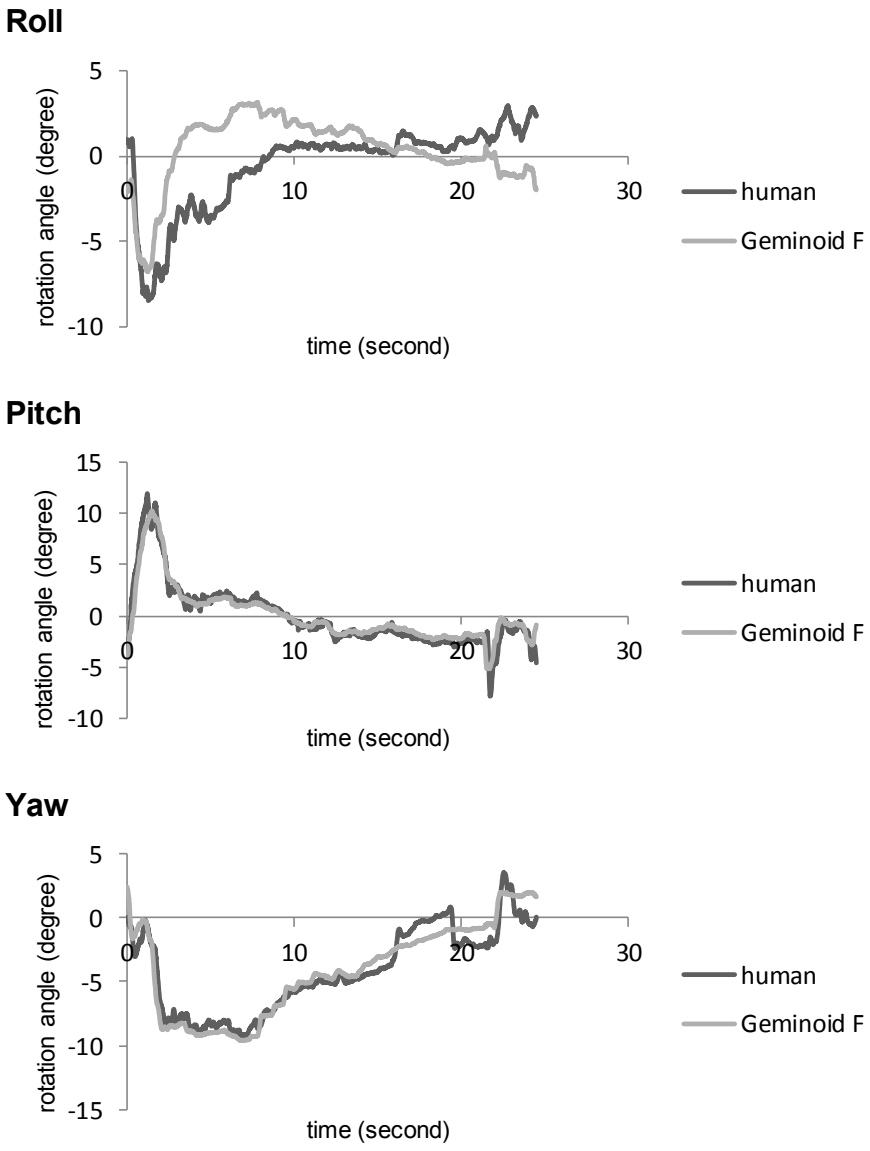


Figure 3.12. Motion data from human and Geminoid F.

3.7. DISCUSSION

The effect observed from adding gaze to the developed generation models was not as great as we had anticipated. We speculate that two causes may exist. First, a person may communicate that they are thinking in different ways: a person may look to one side as if recalling something, tilt their head to one side, or do both. We think that doing both is not necessarily perceived as more natural, because it may convey only the same information (i.e., the person is thinking). Also, robots tend to be less complicated in terms of their communicative capabilities than people; due to this perception, it may be acceptable if some information (such as eye movements) which is normally present with humans is abstracted, or not present for robots. This highlights another interesting possibility; by removing extraneous, misleading or undesired communicative cues which can be observed in humans (such as involuntary twitching), robots could one day become capable of conveying information more clearly than actual humans; thereby, contributing much to service tasks in which such natural communication is an asset; e.g. teaching or theatre performances.

This study focused on exploring how generated nodding, head tilting, and gaze may contribute to the perceived naturalness of a communicating robot. Evaluation was performed both with video footage and with having participants converse face to face with the actual robots, during which three different humanoid robots were used.

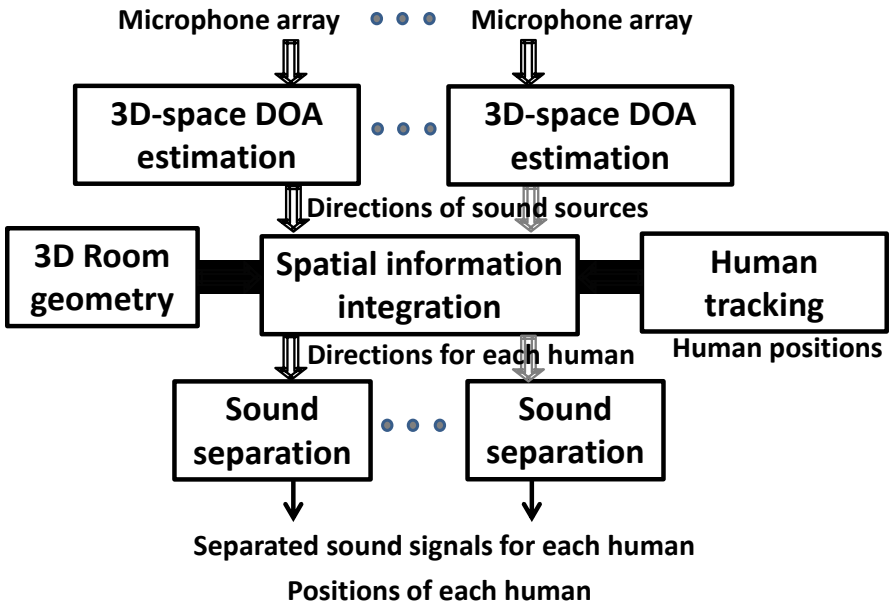
However, the generality of the current work is necessarily limited by target group, choice of robot, cues for generating head motions, and the investigated motions. In this study, Japanese of various ages evaluated our robots for naturalness, but it remains to be shown if results will be exactly the same in different countries or for different cultures. Likewise, it would be interesting to investigate mechanisms for increasing perception of naturalness in non-humanoid communicative robots. Last, we can imagine that other

motions during a conversation, such as posture shifts, or hand motions, could also have an interesting effect on naturalness.

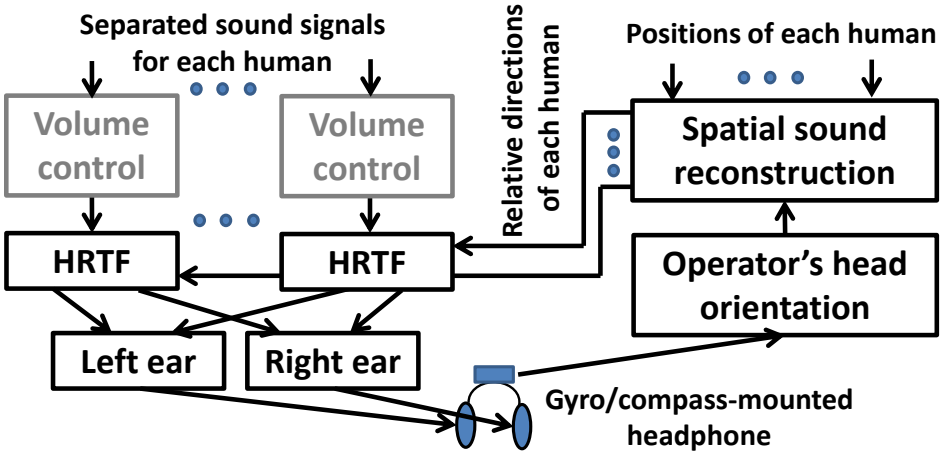
4. AUDITORY SCENE MANIPULATION FOR TELE-OPERATOR

4.1. DESCRIPTION OF THE PROPOSED TELE-PRESENCE SYSTEM

The proposed system can be split in two parts, one for the remote robot environment side, and the other for the tele-operator side. The robot side corresponds to decomposition of individual sound sources with corresponding positions in the 3D space, while the operator side corresponds to synthesis and manipulation of the auditory scene matched with the operator's motions. The block diagrams of both sides are shown in Figure 4.1.



(a) Robot environment side.



(b) Tele-operator side.

Figure 4.1. Block diagram of the proposed tele-presence system.

In the first step for the robot side, the directions of arrival (DOA) of multiple sound sources are estimated in 3D space by multiple microphone arrays. Then, given the 3D room geometry information, the sound directions from all arrays are superimposed in the 3D space with the human positions provided by the human tracking. After integration

of spatial information, the directions of the voice sources for each human are sent to the subsequent sound separation processing. Thanks to the human position information, uninterruptible estimations of the voice source positions can be used for improving separation performance.

In the next step, the voice sources of each human in the remote robot environment are separated by emphasizing the microphone array signals in the target voice directions provided by the 3D sound localization. The resulting separated sound signals and positions for each human are sent to the operator side.

In the operator side, the separated voices are filtered using Head Related Transfer Function (HRTF) database. In this work, a headphone is used for playback, therefore, head orientation tracking is required in order to reduce front-back errors (which is a weakness of HRTFs) and improving direction perception, and for compensating the operator's head movements for synthesizing auditory scenes.

The operator's head orientation is tracked by using gyro sensor and compass attached at the top of a headphone. Appropriate HRTFs are selected according to relative directions of the sound sources and the operator's head orientation for both ears from the HRTF database. The output signals of all sound sources are summed up for each ear, and played in a stereo headphone.

Optional volume control can be conducted for each separated signal for augmented auditory scene manipulation. This option will be described in Section 4.4, by comparing different types of user interfaces for controlling the volume.

In the following sub-sections, we provide details about the system parts.

4.1.1. 3D-SPACE SOUND LOCALIZATION

The directions of arrival (DOA) of multiple sound sources are firstly estimated by multiple microphone arrays, and all sound source directions from each array are

integrated along with the human position information for estimating the location of the sound (voice) sources in the 3D space.

4.1.1.1.3D-SPACE SOUND LOCALIZATION

Locating sound sources in real-world environments is a problem that has been extensively studied so far. The DOA estimation in the present system is based on the MUltiple SIgnal Classification (MUSIC) algorithm, which is a widely used localization algorithm in array signal processing [36]. This method is based on subspace analysis, and is able to estimate directions of multiple sources with high angle resolution. Detail is described in Appendix 4.

The MUSIC method has a constraint that the number of active sources at the moment has to be provided beforehand. Regarding this issue, we adopted the solution proposed in [37], where the number of sources is fixed (considering that the estimation of the number of sources is not robust), and peaks in the MUSIC spectrum exceeding a threshold are picked.

The implemented system can provide 3D-space DOA (i.e., both azimuth and elevation angles) of each sound source with 1 degree resolution. We configured parameters of MUSIC algorithm in order to achieve real-time processing on a single 2GHz CPU core. The DOA estimation is performed every 100 ms blocks. Frame length (i.e., FFT points, which affect FFT performance significantly) has been set to 64 samples, regardless the common setting of 512-1024 samples for reducing computational cost. Regarding the DOA resolution, we used a 2-step approximation method. First, searching directions are restricted to a spherical mesh. In this spherical mesh, every direction has a regular interval of 5 degrees. Then spherical mesh with a step of 1 degree were constructed around the picked directions corresponding to significant MUSIC peaks in first step, for further searching.

4.1.1.2.3D LOCALIZATION OF VOICE SOURCES

In a tele-communication system, the most important sound source that the system should not miss is the human voice. A human tracking system was used in the proposed system to make sure that human positions are tracked continuously even when they are not speaking. In the present work, we employed a 2D human tracking system based on multiple 2D laser range finders (LRF) [38], which is able to provide human positions with accuracies around 5~10 cm, each 10~50 ms.

Given the 3D room geometry, including the positions and orientations of the microphone arrays, the sound directions estimated by each microphone array and the human positions obtained from the human tracking system are superimposed in the 3D space. In a previous work, we have developed methods for 3D localization of sounds based on multiple microphone arrays [39]. In the present work, we also make use of human position information for improving voice source localization, as described below.

The directions of the voice sources are then estimated for being used by the separation process. We consider that if the line drawn for a detected sound direction passes close to the mouth position of a detected human, it can be considered as corresponding to a voice activity of that human. However, the human tracking provides position information only in 2D space, so that information about the mouth height (z-coordinate of the voice source position) is missing. We then imposed some constraints on the possible mouth height range from 1.0 to 1.6 m, considering the cases where a human is sitting or standing. Thus, the mouth height is estimated as the middle point in the shortest line between the lines drawn over a sound direction and a human position, if two conditions are satisfied: 1) the distance between the lines is smaller than 30 cm, and 2) the middle point is within the possible mouth height range (1.0 to 1.6 m). In intervals where the above condition is satisfied, the sound directions are sent to the sound separation block. In non-speech intervals or in intervals where the directivity was not

high enough, the last computed mouth height and the current human positions are used to re-compute the directions relative to the array, which are sent to the sound separation block.

4.1.2. SOUND SEPARATION

Parallel sound separation processes are conducted for all selected humans in the robot environment. In each separation process (for each human), the multi-channel signals of the microphone array closest to the human position are processed in a way to emphasize the sound coming from the voice direction.

4.1.2.1. BEAMFORMING

For the sound separation processing in the present work, we adopted the widely used Delay-Sum (DS) beamforming technique [40], due to its robustness and low computational costs. Separation is conducted every 10 ms frames, with frame length of 20 ms. The beamforming was implemented in frequency domain, according to the expression:

$$Y_{\mathrm{DS}}(f) = \sum_{\mathrm{m}=1}^{\mathrm{M}} \frac{\mathrm{a}_{\mathrm{m,target}}(f)}{\mathrm{a}_{\mathrm{m,target}}^{\mathrm{H}}(f)\mathrm{a}_{\mathrm{m,target}}(f)} X(f) \qquad (1)$$

where M is the number of microphones, f is the frequency bin, am,target is the position vector (transfer function representing delay and attenuation in each m-th microphone for a given target position), and X is multi-channel input spectrum, and Y is the separated signal spectrum. See more detail in Appendix 3.

4.1.2.2. POST-FILTERING

The separated signals after the DS beamforming have low separation performance mainly in low frequencies (below 1000Hz) due to constraints in the array geometry. Figure 4.2 shows frequency response of delay and sum beamformer which is steered to

180 degrees. We can observe a sharp peak only for frequencies higher than 2000Hz. As a consequence, high frequency noises are strongly attenuated, while low frequency noise components remain present in the separated signals. The design of a high-pass-like post-filter was then proposed in this work, in order to compensate for the low performance in low frequencies.

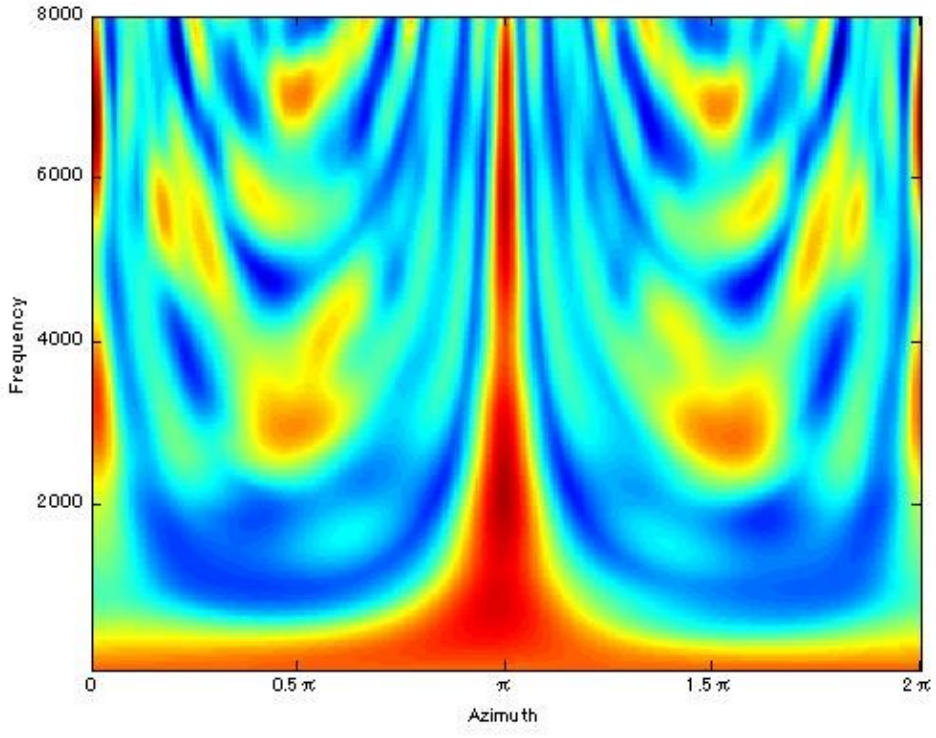


Figure 4.2. Frequency response of delay-sum beamformer (*elevation=30deg*).

Assuming there is a directive target sound S and a non-directive noise N, when the DS beamformer is steered to the signal direction, the output will be following:

$$Y(f) = \boldsymbol{w}_{Sdir}(f) \cdot S(f) + \int_0^{2\pi} (\boldsymbol{w}_{\theta}(f) \cdot N(f)) d\theta \qquad (2)$$

where Yf is the output of DS beamformer at frequency bin f, Sdir is the direction of signal, and wSdir is the beamformer response on direction Sdir. The second term in expression (2) corresponds to the noise leaked into the separated signal. The noise level

depends on the spatial response of the beamformer. To reduce the effects caused by different noise levels in different frequency bands, in this work, we proposed a weight wPF to adjust the amplitude of the beamformer output at frequency f, as follows:

$$w_{PF}(f) = \frac{1}{\int_0^{2\pi} \boldsymbol{w}_\theta(f)\, d\theta} \quad (3)$$

$$Y_{PF} = \sum_f w_{PF}(f) \cdot Y_{DS}(f) \quad (4)$$

where Y_{PF} is the output separated signal after post-filtering.

Further, in order to compensate for the differences in sound levels due to different distances between the microphone array and the voice sources, the amplitude of the output separated signals were normalized according to the following expressions:

$$g_i = \frac{\sum_{n=1}^{N} dist_n - dist_i}{(N-1)\cdot\sum_n^N dist_n} \quad (5)$$

$$Y_i = g_i \cdot Y_{PF,i} \quad (6)$$

where N is the number of separated sources, distn is the distance between the array and the n-th source, gi is the computed normalization factor for the i-th source, and Yi is the separated signal for the i-th source after normalization, to be sent to the tele-operator side.

Finally, the separated sound signals (voices) for each human are sent by TCP/IP in frames of 10 ms length with a header containing time stamps (UNIX time in seconds and milliseconds), and the position of the human (three coordinate axes in millimeters).

4.1.3. HRTF-BASED SYNTHESIS OF AUDITORY SCENES

A common way to synthesize a binaural sound that seems to come from a particular position in space is to filter the original sound signal by a pair of HRTFs for the two ears. The proposed system makes use of an open HRTF database, and synthesizes the

auditory scene by compensating the operator's motion, as described in the following sub-sections.

4.1.3.1. HRTF DATABASE

In the present work, we make use of the HRTF database of the KEMAR Dummy-Head measured by a research group at MIT media lab [41]. KEMAR (Knowles Electronics Manikin for Acoustic Research) is a standard based on common human anthropomorphic data and designed for making HRTF measurements. The KEMAR HRTF database describes the frequency response that results from a complex interaction of anatomically-based reflection and diffraction effects, in the KEMAR dummy-head for different sound source directions.

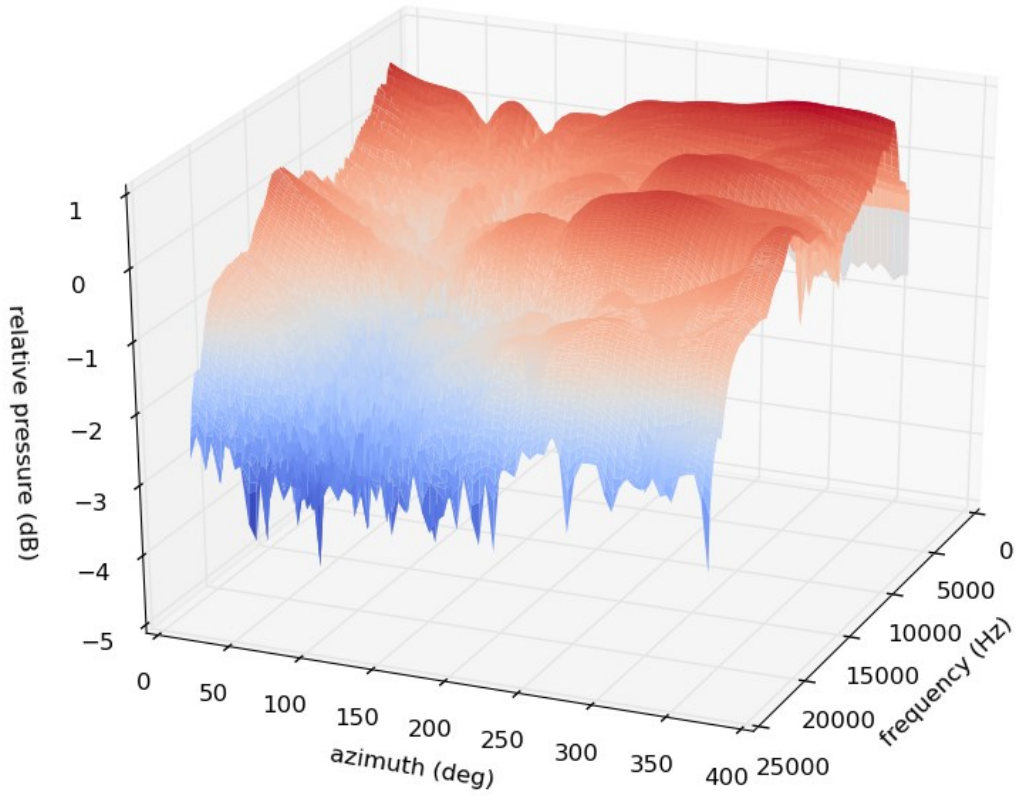


Figure 4.3. Amplitude of KEMAR HRTFs (left ear).

The database includes measurements consisted of head-related impulse response (HRIR, impulse response corresponding to the head related transfer function) for the left and right ears, performed on a KEMAR dummy-head at 710 different positions sampled at elevations from -40 degrees (40 degrees below the horizontal plane) to +90 degrees (directly overhead). Each HRIR was obtained at a sampling rate of 44.1 KHz, and has 512 samples.

Figure 4.3 shows relative sound pressure on the left ear of KEMAR dummy head when sounds come from azimuth plane.

4.1.3.2. HEAD ORIENTATION TRACKING AND SYNTHESIS OF AUDITORY SCENES

It was stated in the introduction that there are two problems when using HRTF to synthesize an auditory scene through headphones, without compensation for the operator's head and body movements. The first is the mismatch in the coordinates of the auditory scenes in the robot side, when the operator's head moves. The second is the front-back confusion.

To deal with these problems, a low latency head orientation tracking is needed. In this work, we attach a gyro sensor and a compass to the top of a headphone to track the rotation of the operator's head. Head angles around the three axes are sent to the system either by serial port or by Bluetooth, each 10~20ms. In following experiments, we used serial connection to lower latency of transmission.

Once the operator's head orientation is received, this angle is subtracted from all voice source directions, to match the auditory scenes with the operator's movements. The system selects proper Head Related Impulse Responses (HRIR) from the HRTF database, corresponding to the matched voice source directions, and convolves each of the separated signals with the corresponding HRIRs. The output signals of all sound sources are summed up for each ear, and played in a stereo headphone.

4.2. SYSTEM EVALUATION

The system evaluation in this section is divided in two parts. In Section 3.1, we first evaluate the effects of allowing head rotation on the perception of auditory scenes by using HRTFs. In Section 3.2, we evaluate the effects of the proposed system during interactions with humans in the robot environment.

4.2.1. HRTF-BASED SYNTHESIS OF AUDITORY SCENES

We first conducted subjective experiments to evaluate the effects of allowing head rotation in sound direction perception, by using HRTFs.

Two separately recorded speech sounds were convolved with HRIRs selected based on two different directions and the orientation of operator's head (as described in Section 4.1.3), and then were played simultaneously to the operator's headphones. The two directions of the sounds were randomly selected from one of the following eight directions around the robot {0, 45, 90, 135, 180, 225, 270 and 315 degrees} (corresponding to North, Northeast, East, Southeast, South, Southwest, West and Northwest.)

Twenty subjects (college students not involved in robotics or acoustic research) participated in this experiment. They were asked to annotate the perceived direction where each sound comes from, given the eight directions above as options. The experiments were conducted for two conditions: allowing head rotation, or keeping the head orientation fixed.

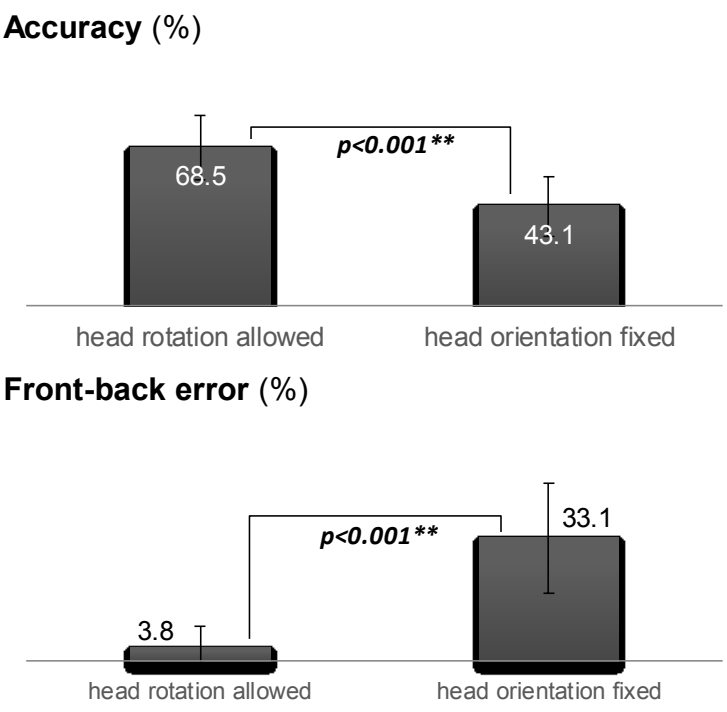


Figure 4.4. Direction perception results for two conditions: "head rotation allowed" and "head orientation fixed".

We compared the accuracy rate of direction perception and front-back error rate, and conducted t-test on these comparisons. The top panel of Figure 4.4 shows means and standard deviations of the resulted accuracy rates. We can observe that the accuracy rate of allowing head rotation was significantly higher than keeping the head orientation fixed ($t=1.70$, $p<0.001$).

The bottom panel of Figure 4.4 shows the results for front-back error rates. It can be observed that the front-back error rate was significantly reduced ($t=-7.55$, $p<0.001$) by allowing head rotation for direction perception.

These results confirm that head rotation improves the perception of direction of sounds through binaural sounds synthesized by HRTFs.

4.2.2. EVALUATION OF THE PROPOSED SYSTEM

Subjective experiments were conducted to evaluate the effectiveness of the proposed system during interactions with humans in the robot environment. As a baseline for comparison, we used binaural microphones located at both sides of the robot's head. These binaural microphones are referred as robot's "ears".

Figure 4.5. External appearance of the Telenoid R3 (top left), operator environment (bottom left) and the robot environment where interaction experiments were conducted (right).

In this experiment, we used a humanoid robot which has two "ears". The Telenoid R3 (top-left picture of Figure 4.5), which is a minimally-designed robot (i.e., it has a neutral face in terms of gender and age), was used in this experiment. It has a head rotatable around three independent axes and two microphones in both sides of head. Thus we can create 3D stereo sound sensation by playing the sound captured by the two

microphones in each ear. For reproducing the head dynamics, operator's head movements are linearly mapped to the robot's head.

The bottom-left picture of Figure 4.5 shows the operator environment, and the right picture of Figure 4.5 shows the robot environment where the interaction experiments were conducted. Red arrows indicate microphone arrays used in this experiment. The 3D sound direction estimation was conducted in the three microphone arrays, indicated by the red arrows in Figure 4.5. The array on the table is 16-channel with the microphones (SONY ecm-c10) arranged over a semi-sphere of 30 cm diameter, while the two attached in the ceilings are 8-channel each, with circular shape with 15 cm diameter. The array on the table was used for sound separation.

Twenty subjects participated in the experiment. The subjects are the same of the previous experiment.

Subjects were asked to talk to one experimenter (a research assistant) in the robot environment side while tele-operating the robot. The experimenter moved to four different directions randomly during the conversation, and subjects had to estimate which direction his conversation partner was in, without visual information. Answer options were restricted to eight directions around the robot position (as in the previous experiment).

Subjective scores were also collected through the following questionnaire, after the conversation. Each question was answered by a 1 to 7 scale where 1 represents worst and 7 stands for best.

- **Perceived degree of sense of presence.**
- **Perceived degree of listenability**

The top panel of Fig. 4 shows means and standard deviations of the accuracy rates for direction perception, for the proposed system and the robot's "ears". A t-test was

conducted and showed a significant difference between the two conditions (t=9.59, p < 0.001).

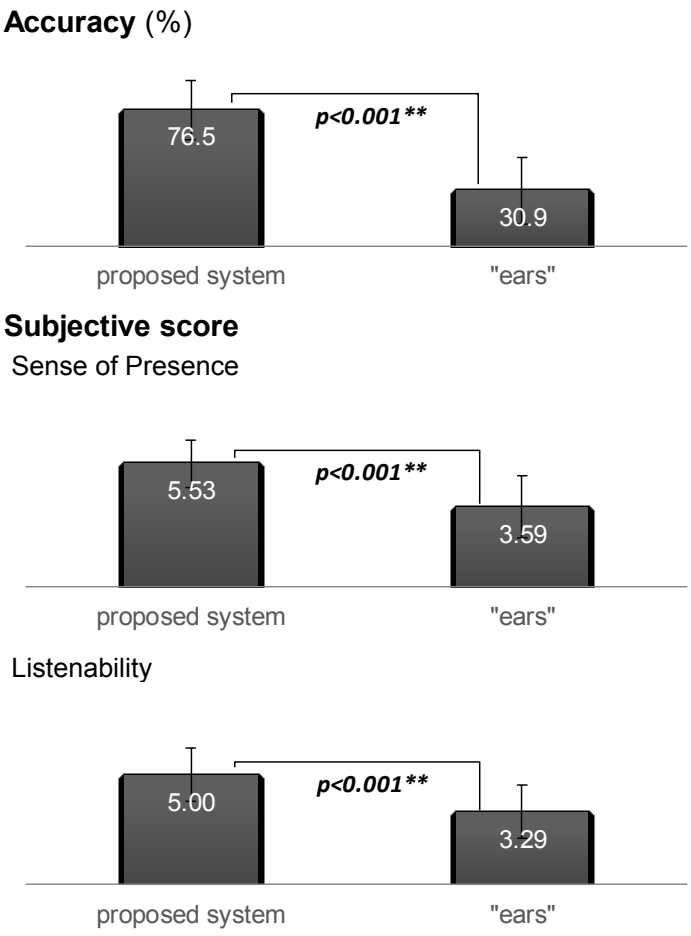


Figure 4.6. Accuracy rates for direction perception and subjective scores (1 to 7 scale) for sense of presence and listenability, in two conditions: "proposed system" and "robot's ears".

Similar trends were found for sense of presence and listenability. The middle and bottom panels of Figure 4.6 show means and standard deviations of subjective scores for the perceived sense of presence and perceived listenability. In both comparisons, the subjective scores for the proposed system were significantly higher than simply using the robot's "ears" (t=6.68, p < 0.001 and t=4.86, p<0.001).

It is easy to understand the differences in perceived listenability between the proposed system and the robot's "ears" since the sound obtained by build-in microphones

are susceptible to motor noise inside the robot and consequently results in a lower signal-to-noise ratio. Regarding the difference for accuracy rate and perceived sense of presence, one possible reason is that the movable range of robot's head is different from human neck. The robot's head could hit its maximum angle while the operator is still rotating his head. This causes a mismatch between the human operator and the robot's head orientation. Another reason is that the "ears" have no earlobes, so that front-back perception is poor.

Participants also reported that sound image only changes its volume but not distance when sound source moved. This problem is discussed in next section.

4.3. USING CALCULATED IMPULSE RESPONSE TO FACILITATE SENSE OF DISTANCE

Regarding the problem that participants could not get proper distance cue during sound sources changed its range in previous experiment, a possible reason could be that HRTFs we used was measured in a fixed range (1.4m). When the sound sources moved, the amplitude of HRTFs is recalculated to fall off in inverse proportion to the distances, and ITD (interaural time difference) remains the same based on plane wave assumption. However, the fact is when the range of sound source changed, both ILD (interaural level difference) and ITD will change as well (ILD increased with decreasing distance because of the decreasing of head scattering effect, ITD decreased with decreasing distance), especially at close distance that curvature of the wave front become significant.

4.3.1. CALCULATION OF IMPULSE RESPONSE

Measuring HRTFs from every single human subject and to cover all range and directions is impracticable. Furthermore, the measurement of near field HRIRs on a human subject is technically difficult [Brungart 2002]. Therefore, we investigated the effect of calculated impulse response in horizontal plane localization. Assuming sounds arrived at ears with spherical wavefront, the response of an impulse signal at range r in frequency domain is given by:

$$a_{(f,r)} = w_r e^{-j2\pi f c^{-1} r} \qquad (7)$$

where f is frequency, c is sound propagation speed in air, r is distance of sound source and ear, w_r is attenuation weight which is inversely proportional to range r.

In a condition that a human subject with head diameter (distance between two ears) of d, for each ear, the range r in equation (7) will be replaced as:

$$r' = ((r\cos\theta \pm d/2)^2 + r\sin\theta^2)^{1/2} \qquad (8)$$

where θ is the direction to sound sources. The impulse response at two ears from location with direction θ and range r can be calculated by evaluating equation (7) and (8).

We used these calculated impulse responses as transfer functions to investigate ITD's effect in sound source localization and range perception.

4.3.2. EVALUATION OF CALCULATED RESPONSE

A subjective experiment is conducted to evaluate the effects of calculated impulse in sound localization and distance feeling.

A recorded speech sound was used for this experiment. This speech sound was convolved with HRIRs selected from HRTF database and IRs calculated based on direction and range of sound source, and played to participants sixteen times (eight times for HRTFs and eight times for calculated IR). The direction of sound was selected randomly as previous experiment described in section 5.2.1 and Range of sound source is changed from 0.5m to 2m randomly.

Twenty-eight subjects participated in this experiment. They were asked to annotate the direction where sound comes from, and keep their head orientation fixed. A questionnaire was also collected for perceiving sense of distance with following question:

- **How do you feel the change of distance instead of volume?**

The question was answered by 1 to 7 scales where 1 represents "only volume changed" and 7 stands for "distance changed clearly".

The top panel of Figure 4.5 shows means and standard deviations of the localization accuracy for using HRTFs and calculated IRs. We can see two conditions have similar results, and slightly higher than previous experiment described in section 3.1

since the task is easier (less sound sources). T-test result also show that there is no significant difference between these two conditions (t=-0.24, p=0.41).

Subjective scores for sense of distance shows difference results (see the bottom panel of Figure 4.7). Scores for using calculated IRs is significantly higher than the use of HRTFs (t=-3.98, p=0.003).

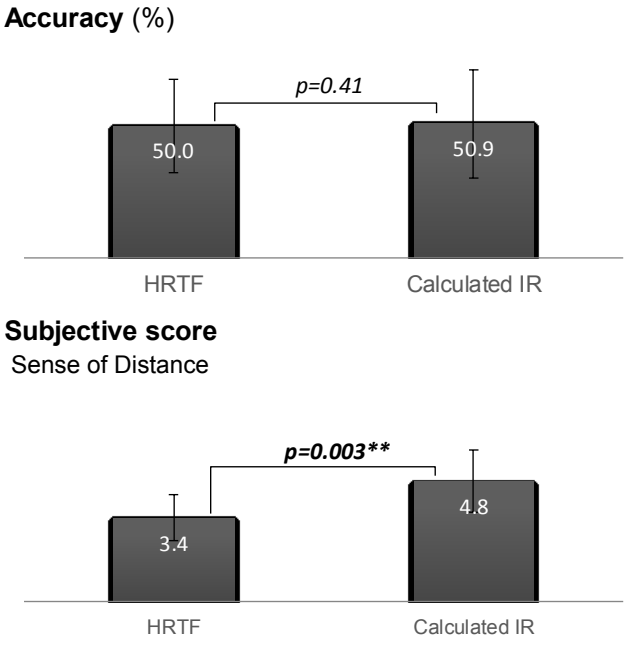


Figure 4.7. Accuracy for direction perception and subjective scores (1 to 7 scale) for sense of distance, in two conditions: "HRTFs" and "calculated IRs".

This result shows that ITD is an important cue for distance perception, and the use of calculated IRs can provide equally accuracy in horizontal plane sound localization. However, because of the absent of ear lobe effect, using calculated IRs cannot help in elevation perception (sound sources with different elevation angles will result same IRs as long as their ranges and azimuth angles are the same).

4.4. AUGMENTED AUDITORY SCENE CONTROL

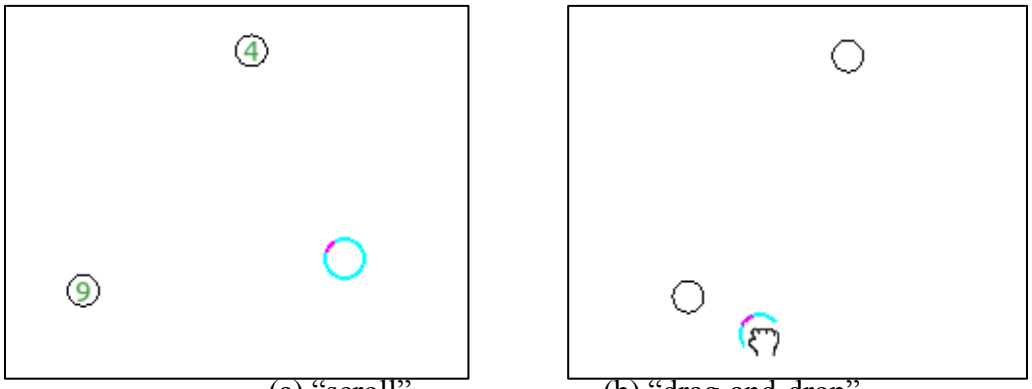
The stereo sound output of the system is a mixed-up of all humans observed and selected by the human tracker. However, there is no guarantee that the voice level of every human is suitable, since for example, some can speak louder than others. In these cases, it is desirable that the operator is able to adjust the sound volumes of each human in the environment. Further, the operator may desire to (virtually) approach to one of the humans, even though the robot cannot (physically) move.

In order to allow the control of 3D auditory effects, including the virtual placement of the operator, three types of user interfaces were designed and evaluated.

4.4.1. DESCRIPTION OF THE PROPOSED USER INTERFACES FOR AUGMENTED AUDITORY SCENE CONTROL

Three types of user interfaces were proposed taking into account different manners to control the manipulation of auditory scenes in the robot's environment.

In the first one, the user can control the sound level of each person (detected by the human tracker) independently. The left panel of Figure 4.6 shows a screen shot of this user interface, where humans are shown as circles and the operator (in this case, the position of the robot) is shown as a circle with a pink arc representing the current head orientation. In this user interface, the operator can adjust the volume of a target person by clicking on the human circle corresponding to that person and scrolling the mouse wheel upward/downward for increasing/decreasing the volume. Double-clicking one of the human circles can maximize the volume of that person and minimize the volume of the others. This user interface is denoted as "scroll". The volume levels for each human are shown inside the human circles, as shown in the Figure 4.6.



(a) "scroll" (b) "drag-and-drop"

Figure 4.8. Screen shots of the displays for different user interfaces.

In the second one, the operator's position can be changed by drag-and-drop the operator's circle to a desired position (virtual robot position). We commonly keep a distance between ourselves and the conversation partner depending on the environment and the social relationship. However, in the present user interface, the user can virtually approach to a target person as to maximize the volume of his/her voice. The sound volumes and sound directions are re-computed according to the new relative positions between the robot and the humans (after drag-and-drop). This user interface is denoted as "drag and drop". The right panel of Figure 4.8 shows a screen shot for this user interface.

In the third one, the sound volumes of the sound sources are adjusted according to the operator's head orientation. It is well-known that head orientation and gaze direction signals individuals' attention, upcoming target or goal [42][43]. In order to free the operator's hands, we employed head orientation as a tool for sound volume adjustment. Sound sources in front of the operator are amplified while opposite ones are attenuated. The amplifier was designed to be linearly related to the angle between the sound source and the operator's head orientation. A maximum attenuation factor of 0.6 was attributed for the opposite direction. This user interface is denoted as "face dir". The display of this user interface is similar to the "scroll" one, as shown in Figure 4.7.

4.4.2. EVALUATION OF THE PROPOSED USER INTERFACES

We conducted subjective experiments to evaluate the three user interfaces. A conventional interface using a monaural microphone was also evaluated for comparison.

Sixteen subjects participated in this experiment. The subjects are the same of the previous experiments, excluding four of them who did not conduct the comparison with the conventional one. With-in participants design was used in this experiment. Subjects were asked to talk to two persons (research assistants) through the tele-presence system. The conversation topics were free. The conversations were separated in four sessions, one for each of the user interfaces (including the conventional one). Each session of conversation lasted about three minutes. After every session, subjects were asked to annotate the perceived degrees of usability, sense of presence, and listenability for each condition. The subjective scores were quantified by a 1 to 7 scale (as in the previous experiments).

Figure 4.9 shows the means and standard deviations of the subjective scores. A repeated measures analysis of variance (ANOVA, Bonferroni's posttest) was conducted for each item. Significant effects were found in "usability" and "sense of presence" ($F(3,13)=10.041$, $p=0.001$ and $F(3,13)=8.777$, $p=0.002$). Regarding listenability, the main effect was not significant ($F(3,13)=3.869$, $p=0.068$).

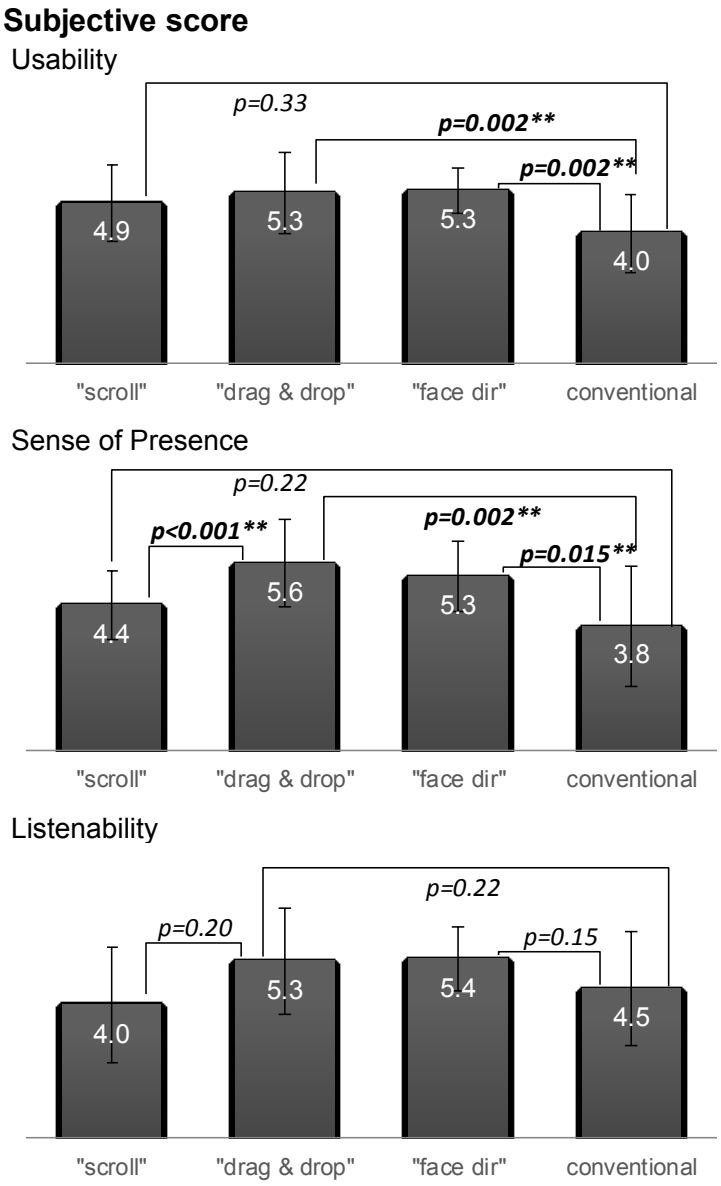


Figure 4.9. Subjective scores (1 to 7 scale) for four types of user interface: "scroll", "drag and drop", "face dir" and "conventional".

Comparing the usability for the four conditions, the subjective scores of two of the proposed user interfaces, "drag and drop" and "face dir", were significantly higher than the conventional one ("drag and drop" vs. "conventional": p=0.002, "face amp" vs. "conventional": p=0.002). No significant difference was observed between condition "scroll" and "conventional" (p=0.33).

Similarly, comparison of sense of presence shows the same trends of usability. Condition “drag & drop” and “face dir” are perceived to provide significantly better sense of presence (“drag and drop” vs. “conventional”: p=0.002, “face dir” vs. “conventional” : p=0.002). Comparison between conditions “scroll” and “conventional” shows no significant difference (P=0.22).

However, regarding the subjective scores for listenalibity, the comparison results show no significant difference between all pairs of conditions (“scroll” vs. “conventional”: p=1, “drag and drop” vs. “conventional”: p=0.22, “face dir” vs. “conventional”: p=0.15).

The results above indicate that the interfaces “drag and drop” and “face dir” are effective user interfaces for tele-presence systems, for improving both usability and sense of presence.

4.4.3. DISCUSSION

The results of the experiment described in this section showed that the subjective scores for both usability and sense of presence for the “scroll” interface were lower than in “drag and drop” and “face dir” interfaces, and showed no significant differences compared to the conventional one. In the interview with the subjects after the experiment, most subjects answered that the operation in interface “scroll” was more complicated than the others. When the operator needs to adjust the sound volume, he/she can drag the circle representing the operator/robot and approach to the circle representing the target sound source using “drag & drop” interface, or simply turn his/her head and face to the target sound sources. However, when using “scroll” interface, the operator needs to first select the target circle and then scroll the mouse wheel to adjust the sound volume, and repeat this process on the other sound sources. This complication is thought to be a potential cause for lower scores in usability and sense of presence.

It is also interesting to note that the subjective scores for sense of presence were the highest for “drag and drop” interface, where the operator virtually changes his/her position, even though the physical robot position was unaltered.

Regarding the listenability scores, although higher mean values were obtained for “drag and drop” and “face dir”, the differences with the conventional one were not judged to be significant. A possible reason is that the number of sound sources in the robot environment was not high, so that the speech contents were listenable even in the conventional system. We guess that the subjective scores would change if more sound sources are in the environment. This is a topic for future investigation.

Finally, regarding the use of the generic HRTFs from a dummy-head in the experiments, possible improvements can still be reached if the HRTFs can be adapted to each user.

5. CONCLUSIONS

The present work proposed a tele-communication robot system that can facilitate tele-presence by generating robot's head motion and manipulating auditory scene for tele-operator.

On robot side, a rule-based nodding and head tilting motion generation model was proposed based on dialogue act function tags, and evaluated in two humanoid robots, for utterances including thinking behaviors. Subjective scores showed that the proposed model including head tilting and nodding can generate head motion with increased naturalness compared to nodding only or directly mapping people's original motions.

Also, as an alternative for utterance motions, a "face up" motion was additionally proposed for robots without movable lips. It was shown that the inclusion of such a motion can improve the perceived naturalness of robots which do not have a mouth, such as the humanoid robot, Robovie R2.

An experiment conducted in which participants conversed with the robots face to face confirmed the validity of the previous video-based evaluation. As well, we found

that where neck motions were properly reproduced in all three axes for the Telenoid that the original motions with gaze movements were perceived as most natural.

Remaining topics for future work include investigating the effects of eye gazing during head motions, and the development of methods for extracting linguistic and paralinguistic information from speech (such as phrase boundaries, disfluencies, and dialogue acts), in order to automate the generation of head motion commands.

On tele-operator side, we proposed and evaluated a tele-presence system which is able to synthesize and manipulate the auditory scene of a communication robot, and increase the auditory sense of presence by the tele-operator.

Subjective experiments were conducted where participants played the role of the tele-operator and talked to others through a remote humanoid robot. It was firstly confirmed that by the reproduction of auditory scenes matched with the operator's head orientation reduced front-back error and improved the perceived localization of sounds. In the interaction experiments with humans in the robot environment, the proposed system was shown to perform with higher accuracy in localization, and higher subjective scores for sense of presence and listenability, compared to the use of binaural microphones attached to robot's head ("ears").

We also investigated the effect of variable ITD (interaural time differences) in distance feeling of near field sound source. An experiment is conducted where perceptual auditory feedback of a recorded speech sounds is synthesized using both calculated impulse response and HRTFs measured from a dummy head. Experimental results showed that compare to HRTFs, calculated impulse response which implied ITD's change in near field can provide higher sense of distance with same accuracy in perceiving direction of sound in horizontal plane.

Three types of user interfaces with ability to control virtual auditory scenes were also proposed and evaluated. Higher subjective scores for usability and sense of presence were obtained by two of the interface types compared to a conventional (monaural) tele-operation system. The two preferred interfaces were: “drag and drop”, which allows the operator to change his/her virtual position in order to amplify the voice of the humans close to the virtual position, and “face dir”, which amplifies the voices in the frontal direction and attenuates the voices from back direction.

APPENDIX

1. LIP MOTION GENERATION

Firstly, formant extraction is conducted on the input speech signal. The first and second peaks are extracted from the Nineteenth order LPC (Linear Predictive Coding) smoothed spectrum which corresponding to the first and second formants of current frame of speech.

Then the formant space given by the first and second formants is transformed to vowel space. The origin of the coordinates in the formants space (denoted as centerF1, centerF2) are adjusted according to speaker-dependent parameters, since the vowel space changes for different speakers due to differences in gender, age and height. Specifically, the new origin is moved to the center of the speaker's vowel space (refer to a subspace of formant space, corresponding to the schwa vowel in English) in the logF1 vs. logF2 space (i.e., the space given by the logarithms of the first and second formants). The next step of speaker normalization is the scaling of the coordinates. This step would be equivalent to a vocal tract length normalization, where the logF1 and logF2 coordinates are stretched or enlarged. Preliminary analysis indicated that scaling factors around 2 for male and 1.8 for female speakers are good approximations. In the present work, the scaling factor is

automatically estimated from center F1, so that values around 450 ~ 500 Hz (average for male speakers) produce scaling factors around 2, and values around 540 ~ 600 produce scaling factors around 1.8.

Finally the axes of logF1-logF2 space are rotated clockwise by about 25 degrees. The new coordinate axes after rotation will be represented by logF1' and logF2'. This rotation process was motivated by the observations that the logF1' axis (after rotation) has good correspondence with the (vertical) aperture of the lips. Figure A1.1 shows examples of distributions of the formant maps for isolated vowels uttered by two speakers (one male and one female), superimposed by the average vowel spaces for Japanese male and female speakers. The new coordinates after translation to the center of the vowel space and rotation are also shown. Note that logF1' values are ordered as /a/ > /e/ > 0 > /o/ ≅ /i/ > /u/, which correspond to the relative lip height variations between the different vowels. The center (schwa) vowel has logF1' = 0.

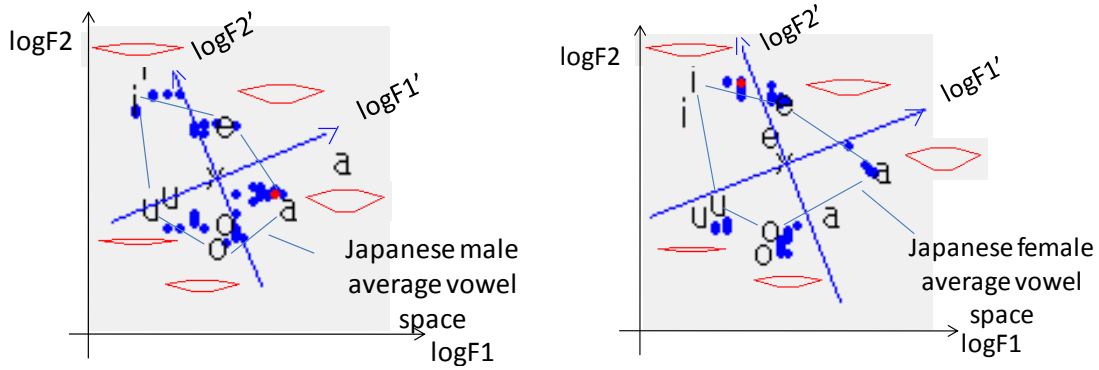


Figure A1.1. Examples of the distributions of single vowel uttered by a male speaker (left), and a female speaker (right).

Normalized lip height values are estimated from the formants as following:

$$lip\_height = 0.6 + height\_scale \cdot logF1'$$

where *lip_height = 0* corresponds to closed mouth, *lip_height = 1* corresponds to a maximally opened mouth, the factor 0.6 corresponds to the aperture for the schwa vowel, and *height_scale* is the scaling factor.

For the lip width, the *F2* value before rotation is used in this work. Although *F2* is more related to the tongue position, there is some relationship with lip rounding. *F2* (or *logF2*) values are ordered like /i/ > /e/ > /a/ > /u/ > /o/, which correspond to the degree of lip spreading in /i/ to lip rounding in /o/. Lip width values can be estimated from the second formant according to the following expression:

$$\Delta lip\_width = width\_scale \cdot \frac{F2 - centerF2}{centerF2}$$

In this equation, Δlip_width gives both positive and negative values. Positive values are obtained when *F2* is higher than *centerF2* (e.g. in /i/ and /e/) where the lips are spread, while negative values are obtained when *F2* is lower than *centerF2* (e.g. in /o/ and /u/) where the lips are rounded. The scaling factor *width_scale* determines the degree of lip spreading/rounding relative to lip height. We used values around 0.5 in this work since they were found to produce human-like lip shapes. Nonetheless, it is important to clarify that the relationship between *F2* and lip width does not follow previous equation strictly for languages with labialized vowels (e.g. in French), and is used here as a first approximation.

Figure A1.1 also shows Representative lip shapes generated for each vowel. In this figure, we can observe that the same vowels of different speakers result similar generated lip shapes, which are benefited by the normalization processing.

The mapping between formants and lip shapes can be constructed in vowel or semivowel intervals. In consonants, where there is a constriction in the vocal tract, the formants are more difficult to be estimated, and its relationship with lip shape is less straight.

In the present work, we use formant range and power constraints for discriminating consonants, as shown in the block diagram of Figure A1.2, and fix a lip height of 0.35, corresponding to an average aperture in consonants. If the low-power interval exceeds a threshold (200 ms), it is judged to be a non-speech interval, and the mouth is gradually closed by a multiplying factor of 0.9, so that the mouth is totally closed after 200 ~ 300 ms.

Generated time series of lip height and width sequences are filtered by a moving average smoothing filter with a window length of 9 (averaging 9 points of input lip shape, including current point, 4 past and 4 future points, with intervals of 10 ms between two points) before conducting actuator command.

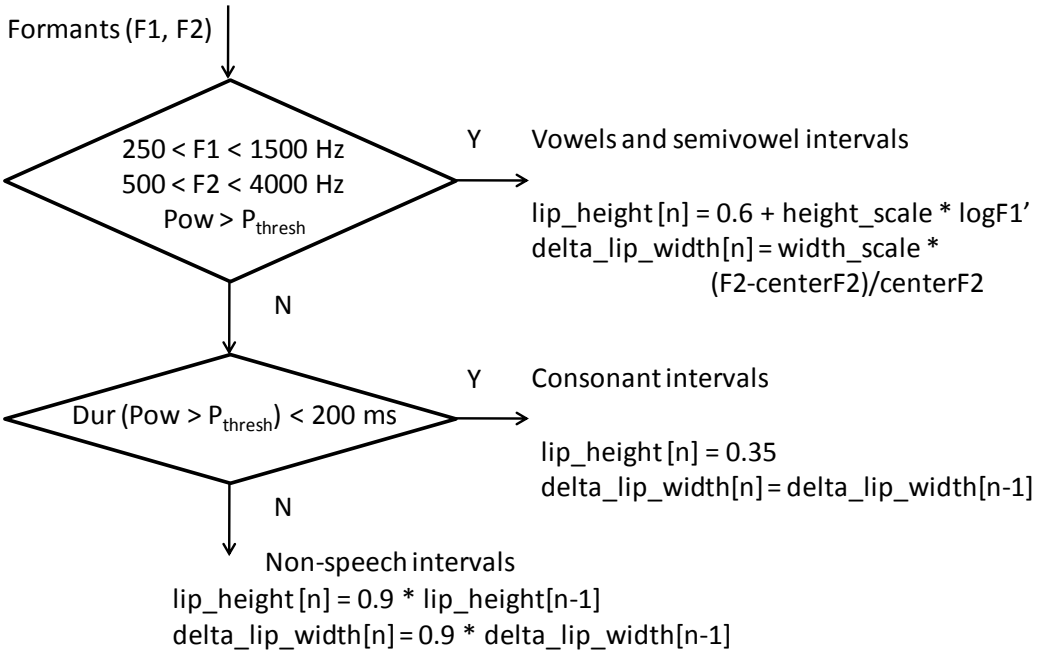


Figure A1.2. Lip motion generation based on formants

The generated actuator commands, obtained in 10ms frame intervals, and the audio packets are sent by TCP/IP to the robot side.

2. ARRAY MANIFOLD VECTORS

Considering a sensor array system includes M sensors, a signal wave from a source in space point can be obtained by each sensor with different phases. Relative positions of signal source with each sensor differ, and result a different propagating route from signal source to each sensor. This difference causes various time delay that signal wave arrive at each sensor. If we denote time delay on each sensor as τ_i, the observations of sensor array can be written as:

$$z(t) = \begin{bmatrix} z_1(t) \\ \vdots \\ z_M(t) \end{bmatrix} = \begin{bmatrix} s(t-\tau_1) \\ \vdots \\ s(t-\tau_M) \end{bmatrix}$$

where *t* is time, $z_m(t)$ is the observation of *m*th sensor, $s(t)$ denotes original signal at time *t*. The observations in frequency-domain, which is Fourier Transform of observations in time-domain, can be described as:

$$Z(\omega) = \begin{bmatrix} Z_1(\omega) \\ \vdots \\ Z_M(\omega) \end{bmatrix}$$

where:

$$Z_m(\omega) = \int_{-\infty}^{\infty} Z_m(t) e^{-j\omega t} dt$$

$$= \int_{-\infty}^{\infty} s(t-\tau_m) e^{-j\omega t} dt$$

$$= e^{-j\omega\tau_m} S(\omega)$$

$S(\omega)$ is the Fourier Transform of original signal as:

$$S(\omega) = \int_{-\infty}^{\infty} s(t)\, e^{-j\omega t} dt$$

We can define a vector as:

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_M \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} e^{-j\omega\tau_1} \\ \vdots \\ e^{-j\omega\tau_M} \end{bmatrix}$$

then the observations in frequency-domain can be denoted as:

$$\boldsymbol{Z}(\omega) = \boldsymbol{aS}(\omega)$$

The defined vector a is called array manifold vector. We used this array manifold vector in source localization and separation process which are described in following sections.

3. DELAY AND SUM BEAMFORMING

Delay and sum beamforming, as a spatial filtering method, is one of the simplest and robust methods for emphasizing directional signal on the basis of difference of physical location of sources. In our microphone array involved environment, the desired signal is speech voice uttered from a speaker's mouth. Observations from microphones are mixture of desired signal and other interfering signals such as other speakers and environment reverberation. Since typically the interfering signals are originated from a spatial point which different from desired speaker's mouth, spatial filtering method can be very useful.

There are two general beamforming methods: time-domain beamforming and frequency-domain beamforming. In time-domain beamforming, a FIR filter is used on each microphone. The beamformer output is formed by a sum of each FIR filtered signal.

$$y(t) = \sum_{m=1}^{M} w_m(t) \otimes z_m(t)$$

Where $z_m(t)$ is the observation from mth microphone, $w_m(t)$ is the filter for mth microphone, $y(t)$ is the output of beamformer, "⊗" notate convolution opeation.

In frequency-domain method, the signal observed from each microphone is separated into narrow-band frequency bins, which can be done with band-pass filter or discrete Fourier Transform, and process the data in each frequency bin separately.

$$Y(\omega) = \sum_{m=1}^{M} W_m^*(\omega)\, Z_m(\omega)$$

Where $Z_m(\omega)$, $W_m^*(\omega)$ and $Y(\omega)$ are Fourier Transformation from $z_m(t)$, $w_m(t)$ and $y(t)$. This equation can also be written in vector notation as:

$$Y(\omega) = W^H(\omega) Z(\omega)$$

$$Z(\omega) = [Z_1(\omega), \cdots, Z_M(\omega)]^T$$

$$W(\omega) = [W_1(\omega), \cdots, W_M(\omega)]^T$$

Figure A3.1 shows diagrams of time-domain and frequency-domain beamforming.

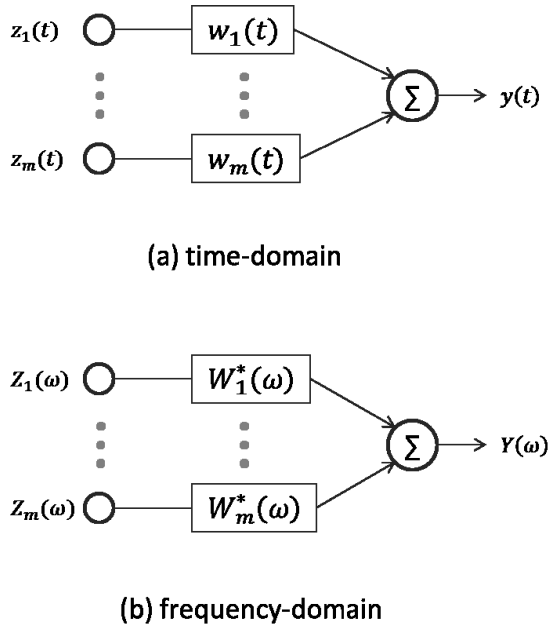


Figure A3.1. Time-domain and frequency-domain beamforming

In our application, plane wave-fronts assumption (i.e. the signal sources are located sufficiently far away from the microphone array that the wave-fronts impinging on the microphones can be seem as plane) is applied for simplified the analysis. A signal from a given direction arriving at each microphone will be differ in time. This difference depends on locations of microphones and direction of incoming signal. A filter $w_m(t)$ or $W_m^*(\omega)$ designed for this specified direction will compensate time difference of each microphones. Regarding this specified directional signal, filtered signals of all microphones will have their phase shifted to the same. Therefore, by summing all filtered signals of each microphone, the signal originated from this specified direction will achieved the maximum output amplitude.

This kind of filter in time-domain can be expressed as:

$$w_m(t) = \frac{1}{M}\delta(t + \tau_m)$$

where τ_m indicate the time delay on *m*th microphone, $\delta()$ is Dirac delta function.

Figure A3.2 shows a processing image of a typical delay and sum beamfoming.

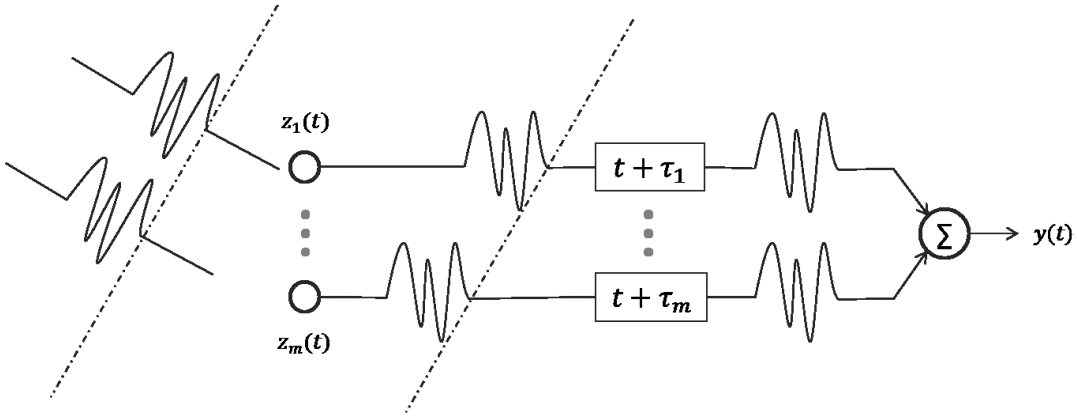


Figure A3.2 Delay and sum beamforming

Filter of delay and sum beamformer in frequency-domain, which refer to Fourier Transform of time-domain filter $w_m(t)$, can be written as:

$$W^H = \frac{1}{M}[e^{j\omega\tau_1}, \cdots, e^{j\omega\tau_M}]$$

If we recall the expression of observation described in Appendix 2:

$$Z(\omega) = aS(\omega)$$

where $\boldsymbol{a}$ denotes array manifold vector:

$$a = [e^{j\omega\tau_1}, \cdots, e^{j\omega\tau_M}]^T$$

We can find that:

$$W = \frac{1}{M}a$$

Combining previous equations, desired signal from given direction can be recovered by:

$$Y(\omega) = W^H aS(\omega) = S(\omega)$$

As well as phases, the amplitude of beamforming output can also be normalized using array manifold vectors:

$$W_{DS} = \frac{1}{a^H a} a$$

The output of delay and sum beamforming can be given as:

$$Y(\omega) = W_{DS}^H Z(\omega)$$

4. MULTIPLE SIGNAL CLASSIFICATION

Determining locations of sound sources is one of important possibilities of a microphone array application. Multiple signal classification (MUSIC) is a method for estimating signal's direction of arrival (DOA), which is a subspace-based spectral method. The MUSIC method has gained significant attention and popularity due to its high resolution and inexpensive computational cost compared to multidimensional search methods.

Considering a microphone array includes M microphones, its observation (the output of all M sensors) will be a matrix sized as M×t, where t is the frame length. In frequency domain, this observation can be written as:

$$\boldsymbol{Z}(\omega) = \boldsymbol{A}(\theta)\boldsymbol{S}(\omega) + \boldsymbol{n}(\omega)$$

where θ is the DOA of impinging signal, $\boldsymbol{A}(\theta)$ is array manifold vector of direction θ which introduced in appendix 2, $\boldsymbol{S}(\omega)$ and $\boldsymbol{n}(\omega)$ denote signal and noise in frequency domain (Fourier transform of time domain signal and noise). We defined the covariance of observation $\boldsymbol{Z}(\omega)$ as observation space:

$$\boldsymbol{R} = E[\boldsymbol{Z}(\omega)\boldsymbol{Z}(\omega)^H]$$

$$= E[\boldsymbol{A}(\theta)\boldsymbol{\Gamma}_s\boldsymbol{A}(\theta)^H + \sigma\boldsymbol{R}_n]$$

where $\boldsymbol{\Gamma_s}$ is signal covariance matrix. If the sources are non-coherent, $\boldsymbol{\Gamma_s}$ will be a full rank matrix. $\boldsymbol{R}_n$ is normalized noise covariance matrix. Assuming noise is zero-mean Gaussian, we have:

$$tr(\boldsymbol{R}_n) = M$$

By doing Eigen decomposition on observation matrix $\boldsymbol{R}$, the eigenvalues $\{\lambda_1, \lambda_2, \cdots, \lambda_M\}$ and eigenvectors $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_M\}$ can be obtained such that:

$$\boldsymbol{R}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i$$

We defined a diagonal matrix $\boldsymbol{\Lambda}$ that all its diagonal entries are eigenvalues of observation matrix:

$$\boldsymbol{\Lambda} = diag[\lambda_1, \lambda_2, \cdots, \lambda_M]$$

and a matrix $\mathbf{E}$ comprised by eigenvectors of observation matrix:

$$\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_M\}$$

Using matrix $\mathbf{E}$, observation space can be transformed into Eigen space by applying Karhunen–Loève theorem:

$$\boldsymbol{y} = \boldsymbol{E}^H \boldsymbol{z}$$

$$= [\boldsymbol{e}_1^H \boldsymbol{z}, \boldsymbol{e}_2^H \boldsymbol{z}, \cdots, \boldsymbol{e}_M^H \boldsymbol{z}]^T$$

Next, we discuss subspace method in three conditions: without noise, with white noise, and with non-white noise.

In an ideal environment without any noise ($\boldsymbol{R}_n = \mathbf{0}$), the observation can be denoted as:

$$\boldsymbol{z} = \boldsymbol{z}_S = \boldsymbol{A}\boldsymbol{s} = \sum_{i=1}^{N} \boldsymbol{a}_i \boldsymbol{s}_i$$

$$\boldsymbol{R} = \boldsymbol{R}_S = \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^H$$

N in previous equation is the number of sound sources, in this case, if we sort eigenvalues $\boldsymbol{\Lambda}$, we will have:

$$\lambda_{N+1} = \lambda_{N+2} = \cdots = \lambda_M = 0$$

Since covariance matrix $\boldsymbol{R}_s$ is Hermitian matrix ($\boldsymbol{R}_s = \boldsymbol{R}_s^H$), we have following equation:

$$\boldsymbol{e}_i^H \boldsymbol{R}_s \boldsymbol{e}_i = \lambda_i$$

When *i* larger than number of sources N, we have:

$$\boldsymbol{e}_i^H \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^H \boldsymbol{e}_i = (\boldsymbol{A}^H \boldsymbol{e}_i)^H \boldsymbol{\Gamma}(\boldsymbol{A}^H \boldsymbol{e}_i) = 0, \qquad i = N+1, \cdots, M$$

Covariance matrix Γ is positive definite matrix, thus:

$$\boldsymbol{A}^H \boldsymbol{e}_i = 0, \qquad i = N+1, \cdots, M$$

$$\boldsymbol{a}_j^H \boldsymbol{e}_i = 0, \qquad i = N+1, \cdots, M, \qquad j = 1, \cdots, N$$

These equations show orthogonality of subspaces. We call $\mathrm{R}(\boldsymbol{A}) = span(\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N)$ signal subspace, the orthogonal complement of signal subspace is noise subspace $\mathrm{N}(\boldsymbol{A}^H) = span(\boldsymbol{e}_{N+1}, \cdots, \boldsymbol{e}_M)$. This relationship can be denoted as:

$$span(\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N) = span(\boldsymbol{e}_{N+1}, \cdots, \boldsymbol{e}_M)^{\perp}$$

We also have:

$$span(\boldsymbol{e}_1, \cdots, \boldsymbol{e}_N) = span(\boldsymbol{e}_{N+1}, \cdots, \boldsymbol{e}_M)^{\perp}$$

since eigenvectors of a Hermitian matrix are orthogonal. Signal subspace $\mathrm{R}(\boldsymbol{A})$ can also be written as:

$$\mathrm{R}(\boldsymbol{A}) = span(\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N) = span(\boldsymbol{e}_1, \cdots, \boldsymbol{e}_N)$$

Conclude it, the non-zero eigenvalues $\{\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \cdots, \boldsymbol{\lambda}_N\}$ and eigenvectors $\{\boldsymbol{e}_1, \boldsymbol{e_2}, \cdots, \boldsymbol{e}_N\}$ are orthonormal basis of signal subspace, while $\{\boldsymbol{e}_{N+1}, \cdots, \boldsymbol{e}_M\}$ comprise noise subspace.

In an environment with white noise, the observation and its covariance become:

$$\boldsymbol{z} = \boldsymbol{z}_s + \boldsymbol{n}_w = \boldsymbol{As} + \boldsymbol{n}_w$$

$$\boldsymbol{R} = \boldsymbol{R}_s + \sigma \boldsymbol{I} = \boldsymbol{A\Gamma A^H} + \sigma \boldsymbol{I}$$

The eigenvalues $\{\lambda_1, \lambda_2, \cdots, \lambda_N\}$ of observation space **R** can be obtained as:

$$\lambda_i = \begin{cases} \lambda_i' + \sigma, & i = 1, \cdots, N \\ \sigma, & i = N+1, \cdots, M \end{cases}$$

where $\boldsymbol{\lambda_i'}$ is eigenvalues of signal covariance matrix. Similarly, we have:

$$\boldsymbol{e}_i^H(\boldsymbol{A\Gamma A}^H + \sigma \boldsymbol{I})\boldsymbol{e}_i = \boldsymbol{e}_i^H \boldsymbol{A\Gamma A}^H \boldsymbol{e}_i + \sigma$$

$$= \sigma, \qquad i = N+1, \cdots, M$$

Signal subspace $\mathrm{R}(\boldsymbol{A}) = span(\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N)$ and noise subspace $\mathrm{N}(\boldsymbol{A}^H) = span(\boldsymbol{e}_{N+1}, \cdots, \boldsymbol{e}_M)$ are the same as in clean environment.

In reality, observations are a mixture of signals and non-white noise which can be denoted as:

$$\boldsymbol{z} = \boldsymbol{z}_s + \boldsymbol{n} = \boldsymbol{As} + \boldsymbol{n}$$

$$\boldsymbol{R} = \boldsymbol{R}_s + \boldsymbol{R}_n = \boldsymbol{A\Gamma A}^H + \boldsymbol{R}_n$$

where noise covariance $\boldsymbol{R}_n = E[\boldsymbol{nn}^H]$ is not diagonal matrix. In this case, generalized Eigen decomposition can be used for whitening noise.

$$\boldsymbol{Re}_i = \lambda_i \boldsymbol{R}_n \boldsymbol{e}_i, \qquad i = 1, \cdots, M$$

We can define a matrix Φ, such that:

$$(\boldsymbol{\Phi}^{-H} \boldsymbol{R} \boldsymbol{\Phi}^{-1}) \boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$$

$$\boldsymbol{\Phi}^H \boldsymbol{\Phi} = \boldsymbol{R}_n$$

$\boldsymbol{f}_i$ is new eigenvectors that $\boldsymbol{f}_i = \boldsymbol{\Phi e}_i$, λ_i is eigenvalues as previous. Replace $\boldsymbol{R}$ with $\boldsymbol{A\Gamma A}^H + \boldsymbol{R}_n$, the equation becomes:

$$\boldsymbol{\Phi}^{-H} (\boldsymbol{A\Gamma A}^H + \boldsymbol{R}_n) \boldsymbol{\Phi}^{-1} \boldsymbol{f}_i$$

$$= (\boldsymbol{\Phi}^{-H} \boldsymbol{A\Gamma A}^H \boldsymbol{\Phi}^{-1} + \boldsymbol{I}) \boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$$

$rank(\boldsymbol{\Phi^{-H} A\Gamma A^H \Phi^{-1}}) = N$ because $rank(\boldsymbol{A\Gamma A^H}) = N$. Like the condition with white noise, the eigenvalues can be written as:

$$\lambda_i = \begin{cases} \mu_i' + 1, & i = 1, \cdots, N \\ 1, & i = N+1, \cdots, M \end{cases}$$

where $\boldsymbol{\mu}_i'$ is non-zero eigenvalues of matrix $\boldsymbol{\Phi}^{-H} \boldsymbol{A\Gamma A}^H \boldsymbol{\Phi}^{-1}$, number '1' in the right hand side of the equations stand for normalized whitened noise power.

If multiply $\boldsymbol{f}_i^H$ to the left hand side of previous equation, we get:

$$\boldsymbol{f}_i^H \boldsymbol{\Phi}^{-H} \boldsymbol{A\Gamma A}^H \boldsymbol{\Phi}^{-1} \boldsymbol{f}_i = 0, \qquad i = N+1, \cdots, M$$

This equation will become:

$$\boldsymbol{e}_i^H \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^H \boldsymbol{e}_i = 0, \qquad i = N+1, \cdots, M$$

by applying condition $\boldsymbol{e}_i = \boldsymbol{\Phi}^{-1}\boldsymbol{f}_i$.

This is the same form as condition with white noise. However, it is worth to note that in non-white noise condition, the eigenvectors from generalized Eigen decomposition have no guarantee that they are mutually orthonormal.

As described above, with or without noise, the eigenvectors of the covariance matrix of observations can span two subspaces: signal subspace (spanned by corresponding eigenvectors of largest *N* eigenvalues) and noise subspace (spanned by corresponding eigenvectors of smallest *M-N* eigenvalues).

The MUSIC algorithm has an additional assumption that the number *N* (number of sound sources) is smaller than the number *M* (number of microphones). As mentioned above, the subspace correspond to noise is spanned by eigenvectors corresponding to the smallest M-N eigenvalues. The principle process of the MUSIC algorithm is searching for directional vector $\boldsymbol{a}$ that orthogonal to the noise subspace, which can be described as:

$$argmin_\theta\{|\boldsymbol{a}^H(\theta)\mathbb{N}|^2\} = argmin_\theta \left\{ \sum_{i=n+1}^{M} |\boldsymbol{a}^H(\theta)\boldsymbol{e}_i|^2 \right\}$$

The MUSIC power spectral is defined as:

$$P_{MUSIC}(\theta) = \frac{\|\boldsymbol{a}^H(\theta)\|^2}{\sum_{i=n+1}^{M} |\boldsymbol{a}(\theta)\boldsymbol{e}_i|^2}$$

$$= \frac{\boldsymbol{a}^H(\theta)\boldsymbol{a}(\theta)}{\boldsymbol{a}^H(\theta)\boldsymbol{E}_n\boldsymbol{E}_n^H\boldsymbol{a}(\theta)}$$

where $\|\boldsymbol{a}^H(\theta)\|^2$ is normalize factor,

$$\boldsymbol{E}_n \overset{\text{def}}{=} [\boldsymbol{e}_{N+1}, \cdots, \boldsymbol{e}_M]$$

Finally, the MUSIC algorithm can be described as follow expression:

$$argmax_\theta\{\boldsymbol{P}_{MUSIC}(\theta)\}$$

$$= argmax_\theta \left\{ \frac{\boldsymbol{a}^H(\theta)\boldsymbol{a}(\theta)}{\boldsymbol{a}^H(\theta)\boldsymbol{E}_n\boldsymbol{E}_n^H\boldsymbol{a}(\theta)} \right\}$$

ACKNOWLEDGEMENT

At the end of this thesis, I am using this opportunity to thank everyone who made my Ph.D. project possible and an unforgettable experience.

First and foremost I want to thank my advisor Prof. Hiroshi Ishiguro, for giving me chance to explore such an exciting and challenging topic and the guidance necessary to make headway. He has taught me, both consciously and unconsciously, how good research is done. I appreciate all his contributions of time, suggestions, and funding to make my Ph.D. experience stimulating.

I would like to express my appreciation and thanks to my supervisor in ATR Dr. Carlos Ishi, for his brilliant ideas and suggestions. I am grateful for the excellent examples he has provided as a successful researcher and mentor.

I also wish to express sincere appreciation to my thesis committee members, Prof. Shogo Nishita and Prof. Youji Iiguni, for taking time to review my thesis and giving helpful comments.

I acknowledge my gratitude to my colleagues in ATR, Dr. Takashi Minato, Kyoko Nakanishi, Mika Morita and Megumi Taniguchi, for their help in sensor setting and experimental data collection.

Finally, I must offer profoundest gratitude from deep my heart to my family, especially my wife Yuan, without whose love, support and encouragement, I could never complete this work.

BIBLIOGRAPHY

[1] Nishio, S., Ishiguro, H., Hagita, N. Can a Teleoperated Android Represent Personal Presence? - A Case Study with Children. *Psychologia*, 50(4): 330-342. 2007.

[2] Sumioka, H., Nishio, S., Minato, T., Yamazaki, R., Ishiguro, H. Minimal Human Design Approach for Sonzai-kan Media: Investigation of a Feeling of Human Presence. Cognitive Computation, 2014.

[3] C. Sidner, C. Lee, L.-P. Morency, C. Forlines. 2006. The effect of head-nod recognition in human-robot conversation. Proc. of IEEE/RSJ Human Robot Interaction (HRI 2006), pp. 290-296, 2006.

[4] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. 2007. Head gestures for perceptual interfaces: The role of context in improving recognition. Artificial Intelligence, 171(8-9): 568-585, June 2007.

[5] H.C. Yehia, T. Kuratate, E. Vatikiotis-Bateson. 2002. Linking facial animation, head motion and speech acoustics. J. of Phonetics, Vol. 30, pp. 555-568, 2002.

[6] M.E. Sargin, O. Aran, A. Karpov, F. Ofli, Y. Yasinnik, S. Wilson, E. Erzin, Y. Yemez, A.M. Tekalp. 2006. Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis. Proc IEEE International Conference on Multimedia, 2006.

[7] K.G. Munhall, J.A. Jones, D.E. Callan, T. Kuratate, E. Vatikiotis-Bateson. 2004. Visual prosody and speech intelligibility – Head movement improves auditory speech perception. Psychological Science, Vol. 15, No. 2, pp. 133-137, 2004.

[8] H.P. Graf., E. Cosatto, V. Strom, F.J. Huang. 2002. Visual prosody: Facial movements accompanying speech. Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (FGR'02), 2002.

[9] J. Beskow, B. Granstrom, D. House. 2006. Visual correlates to prominence in several expressive modes. Proc. Interspeech 2006 – ICSLP, pp. 1272-1275, 2006.

[10] C. Busso, Z. Deng, M. Grimm, U. Neumann, S. Narayanan. 2007. Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. IEEE Trans. on Audio, Speech and Language Processing, March 2007.

[11] Y. Iwano, S. Kageyama, E. Morikawa, S. Nakazato, K. Shirai. 1996. Analysis of head movements and its role in spoken dialogue. Proc. International Conference on Spoken Language Processing (ICSLP'96), pp. 2167-2170, 1996.

[12] M.E. Foster and J. Oberlander. 2007. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. Language Resources and Evaluation, 41(3-4), 305-323, Dec. 2007.

[13] Popescu, V. G., Burdea, G. C., Bouzit, M., Hentz, V. R. A virtual-reality-based telerehabilitation system with force feedback. IEEE transactions on Information Technology in Biomedicine. 4(1): 45-51. 2000.

[14] Piron, L., Turolla, A., Agostini, M., Zucconi, C., Cortese, F., Zampolini, M., Zannini, M., Dam, M., Ventura, L., Battauz, M., Tonin, P. Exercises for paretic upper limb after stroke: a combined virtual-reality and telemedicine approach. J. of Rehabilitation Medicine. 41(12): 1016-1020(5). 2009.

[15] Billinghurst, M., Cheok, A., Prince, S., Kato, H. Real world teleconferencing. IEEE Computer Graphics and Applications. 22(6): 11-13. 2002.

[16] Ogi, T., Yamada, T., Tamagawa, K., Kano, M. Immersive telecommunication using stereo video avatar. Proceedings of Ieee Virtual Reality. Yokohama, Japan. 45-51. 2001.

[17] Bullinger, H., Riedel, O., Breining, R. Immersive Projection Technology- Benefits for the Industry, International Immersive Projection Technology Workshop, 13-25, 1997.

[18] Pulkki, V. Spatial sound reproduction with directional audio coding. J. Audio Eng. Soc. 55(6): 503-516. 2007.

[19] Laitinen, M., Kuech, F., Disch, S., Pulkki, V. Reproducing applause-type signals with directional audio coding. J. Audio Eng. Soc. 59(1/2): 29-43. 2011.

[20] Nishiura, T., Yamada, T., Nakamura, S., Shikano, K. Localization of multiple sound sources based on a CSP analysis with a microphone array. Proceeding of ICASSP 2000. Istanbul. II1053-1056 vol.2. 2000.

[21] Khalil, F., Jullien, J. P., Gilloire, A. Microphone array for sound pickup in teleconference systems. J. Audio Eng. Soc. 42(9): 691-700. 1994.

[22] Hamalainen, M., Myllyla, V. Acoustic Echo Cancellation for dynamically steered microphone array system. 2007. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY, USA. 58-61. 2007.

[23] Rumsey, F. Spatial Audio. Focal Press, 2001.

[24] Meyer, E., Neumann, E. Physical and Applied Acoustics: An Introduction. Academic Press, New York, 1972. ISBN 0124931502.

[25] Cheng, C. I., Wakefield, G. H. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. J. Acoust. Soc. Am, 49(4):231-249, April 2001.

[26] Iwaya, Y., Suzuki, Y., Kimura, D. Effects of head movement on front-back error in sound localization. Acoustical Science and Technology. 24(5): 322-324. 2003.

[27] Perret, S., Noble, W. The effect of head rotations on vertical plane sound localization. J. Acoust. Soc. Am. 102(4): 2325-2332. 1997.

[28] M.B. Stegmann, D.D. Gomez, “A brief introduction to statistical shape analysis,” published online, 2002.

[29] C.T Ishi, J. Haas, F.P. Wilbers, H. Ishiguro, and N. Hagita. 2007. Analysis of head motions and speech, and head motion control in an android. Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007), 548-553, 2007.

[30] C.T. Ishi, H. Ishiguro, N. Hagita. 2006. Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts. Proc. Interspeech’2006 - ICSLP, pp. 2006-2009, 2006.

[31] M. DeBoer and A. M. Boxer. 1979. Signal functions of infant facial expression and gaze direction during mother-infant face-to-face play. Child Development, vol. 50, no. 4, pp. 1215–1218, 1979.

[32] S. R. H. Langton, R. J. Watt, and V. Bruce. 2000. Do the eyes have it? Cues to the direction of social attention. Trends in Cognitive Sciences, vol. 4, no. 2, pp. 50 – 59, 2000.

[33] F. Kaplan and V. Hafner. 2004. The challenges of joint attention. Interaction Studies, pp. 67–74, 2004. [Online]. Available: http://cogprints.org/4067/

[34] Y. Nagai, M. Asada, and K. Hosoda. 2006. Learning for joint attention helped by functional development. Advanced Robotics, vol. 20, pp. 1165–1181(17), October 2006.

[35] C. T. Ishi, C. Liu, H. Ishiguro, N. Hagita. 2011. Speech-driven lip motion generation for tele-operated humanoid robots. International Conference on Auditory-Visual Speech Processing, 2011.

[36] Schmidt, R. Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation, 34, 276-280, 1986.

[37] Ishi, C. T., Chatot, O., Ishiguro, H., Hagita, N. Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. In Proceeding of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 09). St. Louis, MO, USA. 2027–2032. 2009.

[38] Glas, D.F. et al, 2007. Laser tracking of human body motion using adaptive shape modeling. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007), 602-608. 2007.

[39] Ishi, C., Even, J., Hagita, N. (2013). Using multiple microphone arrays and reflections for 3D localization of sound sources. In Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013), 3937-3942, Nov., 2013.

[40] Dudgeon, D. E. Fundamentals of digital array processing. Proceedings of the IEEE. 65(6): 898-904. 1977.

[41] Gardner, W. G., Martin, K. D. HRTF measurements of a KEMAR. J. Acoust. Soc. Am. 97(6):3907-3908, Jun. 1995.

[42] Langton, S. R., Watt, R. J., Bruce, I. I. Do the eyes have it? Cues to the direction of social attention. Trends Cog. Sci. 4, 50–59, 2000.

[43] Yokoyama, T., Noguchi, Y. Kita, S. Attentional shifts by gaze direction in voluntary orienting: evidence from a microsaccade study. Exp. Brain Res. 223, 291–300, 2012.

PUBLICATIONS

Journal Papers

石井カルロス寿憲, 劉超然, 石黒浩, 萩田紀博 (2013). 遠隔存在感ロボットのためのフォルマントによる口唇動作生成手法, 日本ロボット学会誌, Vol. 31, No. 4, 83-90, Apr. 2013.

劉超然, 石井カルロス寿憲, 石黒浩, 萩田紀博 (2013). 人型コミュニケーションロボットのための首傾げ生成手法の提案および評価, 人工知能学会論文誌, vol. 28, no. 2, pp. 112-121, January, 2013.

Liu, C., Ishi, C., Ishiguro, H., Hagita, N. (2013). Generation of nodding, head tilting and gazing for human-robot speech interaction. International Journal of Humanoid Robotics (IJHR), vol. 10, no. 1, January, 2013.

Conference Papers (peer reviewed only)

Liu, C., Ishi, C., Ishiguro, H. (2015) "Bringing the Scene Back to the Tele-operator: Auditory Scene Manipulation for Tele-presence Systems" HRI2015, accepted. (Acceptance ratio = 25%)

Ishi, C., Liu, C., Ishiguro, H., Hagita, N. (2012). "Evaluation of formant-based lip motion generation in tele-operated humanoid robots," In IEEE/RSJ International Conference on

Intelligent Robots and Systems (IROS 2012), Vilamoura, Algarve, Portugal, pp. 2377-2382, October, 2012. (Acceptance ratio = 45%)

Ishi, C., Liu, C., Ishiguro, H., Hagita, N. (2012). "Evaluation of a formant-based speech-driven lip motion generation," In 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), Portland, Oregon, pp. P1a.04, September, 2012. (Acceptance ratio = 52%)

Liu, C., Ishi, C., Ishiguro, H., Hagita, N. (2012) "Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction," Proceedings of 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI2012), 285-292. (Acceptance ratio = 25%)

Ishi, C., Liu, C., Ishiguro, H., Hagita, N. (2011). "Speech-driven lip motion generation for tele-operated humanoid robots," Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP2011), 131-135. (Acceptance ratio = 58%)

Ishi, C.T., Liu, C., Ishiguro, H., Hagita, N. (2010). "Head motion during dialogue speech and nod timing control in humanoid robots," Proceedings of 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010), 293-300. (Acceptance ratio = 26%)