9th IEEE-RAS International Conference on Humanoid Robots
December 7-10, 2009 Paris, France

Voice quality manipulation for humanoid robots consistent with their head movements

Takuma Otsuka, Kazuhiro Nakadai, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno

Abstract— **This paper presents voice-quality control of humanoid robots based on a new model of spectral envelope modification corresponding to the vertical head motions, and left-right sound-pressure modulation corresponding to the horizontal head motions. We assume that a pitch-axis rotation, or a vertical head motion, and a yaw-axis rotation, or a horizontal head motion, affect the voice quality independently. Spectral envelope modification model is constructed based on the analysis of human vocalizations. Left-right sound-pressure modulation model is established through the measurements of impulse responses using a pair of microphones. The experiments are carried out using two humanoid robots HRP-2 and Robovie-R2. Experimental results show that our method presents the change in the voice quality derived from pitch-axis head movement in a robot-to-robot dialogue situation when the interval between the robots are 50 cm. It is also confirmed that an observable modulation in the voice quality declines as the distance between the robots becomes large. The voice-cast directionality caused by yaw-axis rotation is observable using our model even when the robots stand as far as 150 cm away.**

I. Introduction

We have an increasing number of chances to have conversations with robots thanks to the development of robots intended to interact with humans, such as ROBISUKE [1], Repliee Q2 [2] or ARMAR II [3]. To realize natural and successful conversations between humans and robots, robots must behave and speak the way humans expect them. For example, robots should face the talker or give back-channel feedback with proper timing. The consistency between the robot's voice quality and its body motion is one of the most especially striking factors in robot speech naturalness. When the robot faces upward, the voice should sound strong and clear; when the robot bends down, the voice should become weak and vague.

Changes in the voice quality corresponding to physical posture or motions is part of paralinguistic information. Speech sounds deliver two kinds of information. One is linguistic and literal meanings of the spoken words. The other is paralinguistic information, which conveys a speaker's state such as their feelings, the internal state of the speaker, and a speaker's physical posture, the external state of the speaker.

T. Otsuka, T. Takahashi, K. Komatani, T. Ogata, H. G. Okuno are with Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan. {ohtsuka, tall, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

K. Nakadai is with Honda Research Institute Japan, Co., Ltd., Wako, Saitama, 351-0114, Japan, and also with Graduate School of Information Science and Engineering, Tokyo Institute of Technology. nakadai@jp.honda-ri.com

Existing studies intended to add paralinguistic information to speech signals focus on physically-independent features such as intonation [4] or emotional aspects [5]. These studies provide spoken dialogue systems with natural speech sounds, and as a result, we find it comfortable to use such systems. However, these kinds of paralinguistic information are insufficient for robots because the changes in voice quality caused by their body movement is ignored.

To achieve the consistency between the direction of speech sounds and the robot's face motion, the direction a voice is cast on the azimuth plane is controlled with an ultrasonic directional loudspeaker attached to the robot's waist [6]. However, this approach encounters several problems. First, to match the direction of the robot's face and its speech signal, the loudspeaker must be embedded in the robot's face. However, this is often difficult because ultrasonic speakers are generally too large compared to ordinary robots' faces. Second, ultrasonic loudspeakers are inappropriate for emitting speech signals. The sound from them has little power in frequency bands less than 500 Hz due to their mechanism [7]. Third, this method only copes with the direction of the voice. The voice from the speaker strikes people as unnatural because this approach presents no change in the voice quality related to the robot's vertical head motion.

This paper presents voice direction control on the azimuth plane using a stereo speaker as well as voice quality control based on a new model of spectral envelope modification corresponding to vertical head motions. The model is constructed through the one-third octave band analysis of human speech sounds.

II. Voice manipulation architecture

This section clarifies the problems in voice quality manipulation and then presents an architecture for the voice control of humanoid robots.

A. Problem Statement

The voice should be "manipulated" using an existing voice synthesizer, instead of synthesizing a voice signal from scratch because existing voice synthesizers successfully generate natural speech signals although they scarcely affect physical motions in the head or torso. To be sure, there is a voice synthesizer simulating vocal tract that is able to take into account these physical motions [8]. However, this method has just generated clear consonants, so, it takes a long time for long words or sentences to be generated. In addition, the manipulation should be applied to any words at a low computational cost because the voice should be

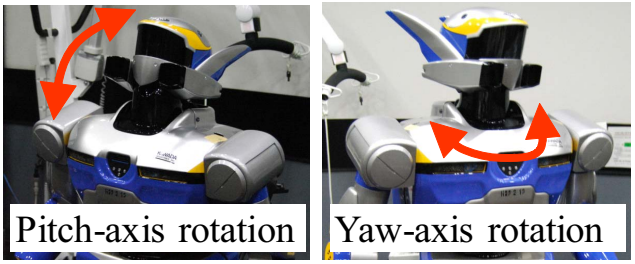

Fig. 1. Head motions in question posed by HRP-2. Pitch-axis on the left and Yaw-axis on the right.

changed before the robot actually speaks after planning what to say. STRAIGHT [9] is one solution to obtain high-quality voice manipulation. However, we have difficulty in utilizing STRAIGHT to manipulate voice qualities for any words because feature points dependent on the phonemes in the spectrogram have to be specified in advance. We use a spectral-envelope control that is applicable to any word generated by existing voice synthesizers.

We focus on correspondence between the voice quality and robot's head motion for the following two reasons. One is that head motions are considered the most relevant of all possible body motions to the changes in voice. The other is that head-motion based voice control is applicable to many humanoid robots because most are able to move their head.

We further divide the head motions into two types: the pitch-axis rotation and the yaw-axis rotation. Figure 1 shows rotations of both axes posed by a humanoid robot HRP-2. The pitch-axis rotation is to nod one's head whereas yaw-axis rotation is to shake one's head right and left. We assume that the pitch movement and the yaw movement affect the voice independently. The pitch rotation changes the spectral envelope of the speech signal because this movement alters the vocal tract, which works as an acoustic filter in accordance with the source filter model [10]. The yaw rotation determines the direction of the voice on the azimuth plane without affecting the vocal tract shape.

Here, the problem statement is specified below.

Input: Original speech signal $x(t)$ and head joint angles, pitch axis θ_p and yaw axis θ_y,
Output: Head-consistent speech signal $\hat{x}(t)$,
Assumption: θ_p and θ_y affect $x(t)$ independently,

where t means time, $x(t)$ and $\hat{x}(t)$ represent speech signals, and θ_p and θ_y are rotation angles of the pitch axis and yaw axis, respectively.

B. Source Filter Model

We assume that the pitch rotation θ_p affects the filter in the source filter model. Figure 2 outlines the source filter model of a human voice [10]. It consists of two parts: sources and a filter. The sources are either pulses generated by the vocal band or a breath noise from the lungs. The filter corresponds to the vocal tract. The filter can be divided into two parts:

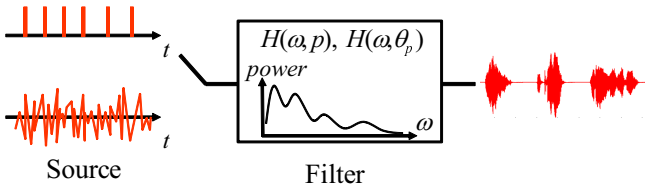

Fig. 2. Source filter model of human voice

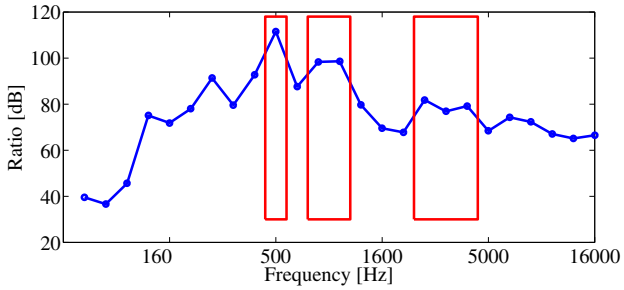

Fig. 3. Power at each frequency band with 0-degree vocalization

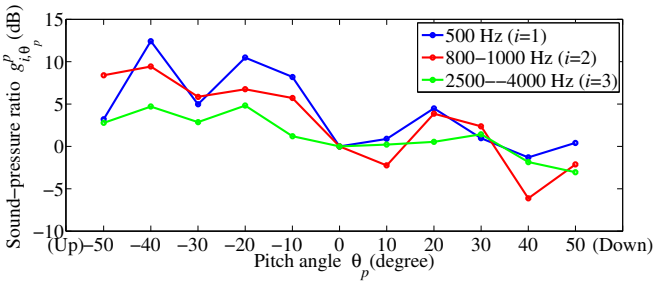

Fig. 4. θ_p–gain model for three bands

$H(\omega, p)$ and $H(\omega, \theta_p)$, where ω indicates a frequency band index and p is a type of phoneme. $H(\omega, p)$ shapes formants or other phonemic features. $H(\omega, \theta_p)$ is the effect of vertical head motions.

1) The construction of $H(\omega, \theta_p)$: We build a model of $H(\omega, \theta_p)$ by inspecting a human voice for various angles θ_p. The recording set up was as follows. Speech signals of a male subject, one of authors, were recorded with a close-talking microphone in an anechoic chamber. A 10-second-long sweep-tone vocalization of vowel [a] was recorded with the subject's head moving 10 degrees at a time from 50 degrees downward to 50 degrees upward. The subject was instructed to vocalize at the same loudness in order to emphasize changes in the spectral envelope without changes in the power. A sweep tone was used to avoid the effect of fundamental frequency and ranged from 261 (Hz) to 523 (Hz), which correspond to musical note C. The recorded voice signal was then analyzed with one-third octave bands, and sound pressure levels for each band compared to the respective levels at 0 degree were calculated.

Figure 3 shows the result of one-third octave band analysis of the 0-degree voice. We choose three frequency bands (500 (Hz), 800 – 1000 (Hz), and 2500 – 4000 (Hz)) to manipulate the voice quality for the following reasons:

1) These bands have more power than other bands.

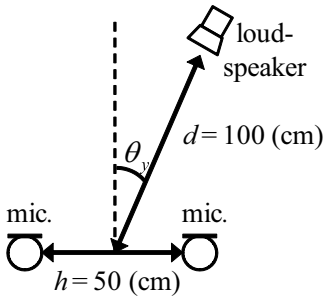

Fig. 5. Setup for a left-right balance measurement

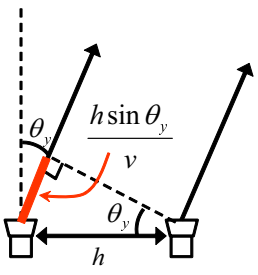

Fig. 6. Geometric illustration of the delay between two loudspeakers

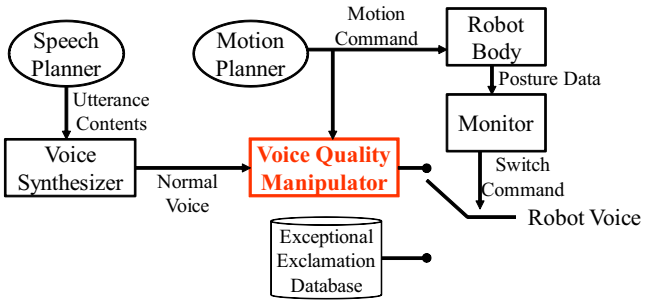

Fig. 8. Voice-manipulation architecture

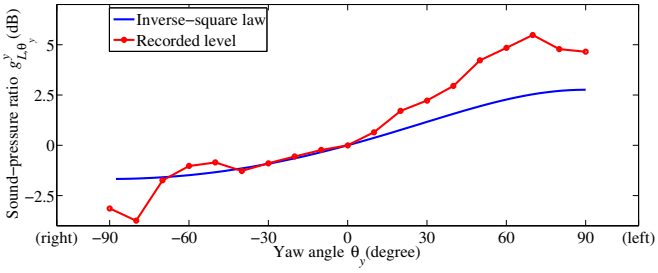

Fig. 7. Empirical and theoretical θ_y–gain model for the left channel

2) These bands also lie in the area where most formants exist.
3) More change in the sound pressure level is observed by varying θ_p.

Therefore, these bands are considered most effective for the voice-quality manipulation. Figure 4 shows the ratios of sound-pressure level g^p_{i,θ_p} in dB to 0-degree voice for each band and θ_p. The upper suffix p means that this is a gain for pitch angle, and i represents the band number. Negative pitch angles indicate facing upward whereas positive ones indicate facing downward.

Observations of vocalizations by another female subject confirm the choice of these frequency bands are valid. Although the sweeping fundamental frequency ranged from 392 (Hz) to 784 (Hz), these three bands have more power than the others. We also confirmed that the gain for each band declined when the subject faced downward. This inclination is observed in our model shown in Figure 4.

C. Azimuth plane control

Here, two methods to present the direction on the azimuth plane are described.

1) A left-right volume balance control of a stereo sound.
2) A delayed stereo sound generation.

We use a pair of stereo loudspeakers and modify the left-right balance of the volume to embody the directional information that yaw-axis head rotation brings about. Embedding a directional loudspeaker in the robot's head seems an attractive method to present the directional information. However, this approach restricts the size of a loudspeaker, and therefore severely deteriorates the sound quality and volume.

We confirmed the left-right balance by measuring impulse responses in an anechoic chamber as shown in Figure 5. The angle θ_y ranged from $-90°$ to $90°$, every $10°$, where $0°$ means the center position and a positive angle means the left direction. On the other hand, a sound pressure theoretically conforms to the inverse-square of the distance from a sound source. The results of both the actual measurement and the simulation of the inverse-square law are shown in Figure 7. The y axis represents the sound-pressure ratio compared to 0-degree sound level. Both plots indicate the left channel. We use the measured result rather than the simulation result because an exaggerated modification is necessary to show the directional information clearly.

Along with the left-right balance control, one channel of the stereo signal is shifted in a time domain to emphasize the directionality. The shift amount is determined geometrically as depicted in Figure 6. The delay time is $h \sin \theta_y / v$, where h is an inter-speaker distance and v denotes sonic velocity.

D. Voice-Manipulation Architecture

This section presents an architecture capable of manipulating the voice quality and coping with unexpected body movements that can cause the voice to be inconsistent with the posture. Figure 8 shows our envisioned voice-manipulation architecture. The voice quality manipulator, shown as a red module in the figure, is presented in this section. The voice manipulation processing proceeds as follows. First, the speech planner determines what to say, and the motion planner determines how to move its body. The utterance content is then sent to the voice synthesizer, which generates a normal voice signal. The motion command is delivered to the robot body for actual movement and to the voice quality manipulator, which changes the voice quality in accordance with the motion command.

Generally, there may be some obstacles to smooth body movements. For example, a robot may stumble, causing an abrupt change in its posture, or something might hit the robot's head, hindering it from moving as planned. If these accidents occur, the voice will be inconsistent with the robot's actual posture. An exceptional exclamation such as "Ouch !" or groan should be produced in such incidents. The exceptional exclamation database in Figure 8 stores such specific voice signals. When the monitor that gathers posture data from the robot's body detects an unexpected

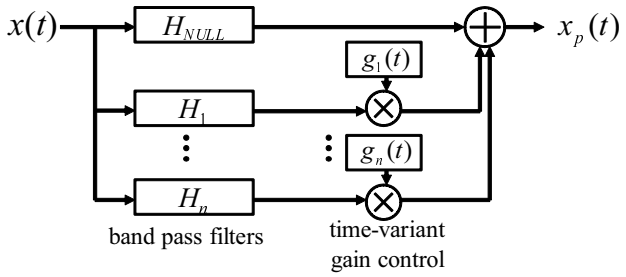

Fig. 9. Flowchart of pitch-axis modification

body movement, the robot speech switches to a specific phrase.

III. Algorithm

This section explains the procedures of our voice manipulation method. The input speech signal is first modulated with a pitch-axis angle, then modulated with a yaw-axis angle. A monaural speech signal $x(t)$ is first modified to $x_p(t)$ with a filter $H(\omega,\theta_p)$. The signal $x_p(t)$ is then doubled into a stereo signal $\hat{\mathbf{x}}(t)=[\hat{x}_L(t)\ \ \hat{x}_R(t)]$, where the left-right balance is controlled with θ_y.

A. Pitch-axis modification

Figure 9 outlines the spectral envelope manipulation corresponding to θ_p. In general, the filter $H(\omega,\theta_p)$ is time-variant because the robot may talk while it is moving its head. If tap weights of $H(\omega,\theta_p)$ are calculated at each sample of $x(t)$ for varying $\theta_p(t)$, it will require a huge computational cost just to calculate the tap weights. However, given that frequency bands to manipulate the sound-pressure level are determined as described in Section II-B.1, we can divide the original signal $x(t)$ into several band-passed signals: $x_{NULL}(t)$ and $x_i(t)$. These signals are in accordance with given frequency bands as shown in Figure 9, where $x_{NULL}(t)$ has no power at all given frequency bands and $x_i(t)$ is the output of the i-th band pass filter. Then, each $x_i(t)$ is amplified by the gain $g_i^p(t)$ in accordance with θ_p. This method requires only computation of $g_i^p(t)$ and reduces the computational cost because the tap weights for each band pass filter are constant.

a) Decomposition: The original signal $x(t)$ is decomposed into $x_{NULL}(t)$ and $x_i(t)$, where $i=1,2,3$. The length of tap weights $h_{NULL}(t)$ and $h_i(t)$ are all 41. The θ_p-invariant component $x_{NULL}(t)$ is calculated as

$$x_{NULL}(t)=(x*h_{NULL})(t), \tag{1}$$

where $*$ represents convolution. The gain of frequency response for h_{NULL} is zero at 500 Hz, 800 – 1000, 2500 – 4000 Hz and one at the other one-third octave band center frequencies up to Nyquist frequency of the signal $x(t)$.

The θ_p-dependent components $x_i(t)$ is calculated as

$$x_i(t)=(x*h_i)(t), \tag{2}$$

where the gain of frequency response for h_i is one at respective frequency bands and zero at the other frequencies. For example, the gain of the second band pass filter is one at 800 and 1000 Hz but zeros at the other center frequencies.

b) Amplification: The gain for each sample $g_i^p(t)$ is obtained by interpolating the gains shown in Figure 4 every 10 degrees using $\theta_p(t)$ as shown in equation (3).

$$g_i^p(t) = \frac{g^p_{i,\theta_m}(\theta_{m+1}-\theta_p(t))+g^p_{i,\theta_{m+1}}(\theta_p(t)-\theta_m)}{10}, \tag{3}$$

$$\theta_{m+1} = (\lfloor\theta_p(t)/10\rfloor+1)\times 10, \tag{4}$$

$$\theta_m = (\lfloor\theta_p(t)/10\rfloor)\times 10, \tag{5}$$

where g^p_{i,θ_m} is the gain of the i-th band corresponding to the angle θ_m in the model. $\lfloor x\rfloor$ is the largest integer equal to or less than x. For example, when $\theta_p(t)=35^\circ$, $\theta_{m+1}=40^\circ$ and $\theta_m=30^\circ$, consequently, $g_i^p(t)=(g^p_{i,30^\circ}+g^p_{i,40^\circ})/2$. The time-variant signals $x_i(t)$ are then amplified as

$$x_{i,g}(t) = x_i(t)\times 10^{\frac{g_i^p(t)}{10}}. \tag{6}$$

Note that $g_i^p(t)$ is in dB. Therefore, it should be transformed into a scale $10^{\frac{g_i^p(t)}{10}}$.

c) Reconstruction: The voice manipulation in accordance with $\theta_p(t)$ is completed by adding up time-invariant component $x_{NULL}(t)$ and time-variant components $x_{i,g}(t)$. Therefore,

$$x_p(t)=x_{NULL}(t)+\sum_{i=1}^{3}x_{i,g}(t). \tag{7}$$

B. Yaw axis modification

The pitch-axis modulated and monaural signal $x_p(t)$ is first doubled to a stereo signal $\mathbf{x}_{ste}(t)$, both of which channels equal $x_p(t)$. Both channels of the stereo signal, $x_L(t)$ and $x_R(t)$, are amplified in accordance with the control model shown in Figure 7. Then, in our implementation, the left channel $x_L(t)$ is shifted in a time domain to emphasize the directionality.

d) Left-right balance control: The gain for each channel $g_j^y(t;\theta_y(t))$ $(j=L,R)$ is obtained by interpolating the discretely measured gains $g^y_{\theta_n}$ for continuous angle $\theta_y(t)$ as in equations (8) and (9).

$$g_L^y(t;\theta_y(t)) = \frac{g^y_{\theta_n}(\theta_{n+1}-\theta_y(t))+g^y_{\theta_{n+1}}(\theta_y(t)-\theta_n)}{10}, \tag{8}$$

$$g_R^y(t;\theta_y(t)) = g_L^y(t;-\theta_y(t)), \tag{9}$$

$$\theta_{n+1} = (\lfloor\theta_y(t)/10\rfloor+1)\times 10, \tag{10}$$

$$\theta_n = (\lfloor\theta_y(t)/10\rfloor)\times 10. \tag{11}$$

The gains of left and right channel are symmetric as expressed in equation (9). In the next step, each channel is amplified by the respective gain as

$$x'_j(t) = x_j(t)\times 10^{\frac{g_j^y(t)}{10}}(j=L,R). \tag{12}$$

e) Time shift: This time shifting finishes the voice manipulation process. The left channel signal precedes the right one by $h\sin\theta_y(t)/v$ second, where h is the distance

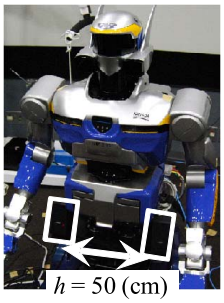

Fig. 10. HRP-2 with its stereo loudspeakers marked by rectangles

Fig. 11. HRP-2 on the left and Robovie-R2 on the right

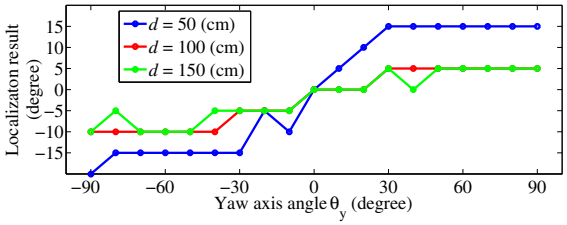

Fig. 12. Sound localization result on the azimuth plane

between two loudspeakers and v is the sonic velocity. The time-shifted signal $\hat{x}_L(t)$ is obtained as follows:

$$\hat{x}_L(t) = x'_L(t - \frac{h \sin \theta_y(t)}{v}), \quad (13)$$

$$\hat{x}_R(t) = x'_R(t). \quad (14)$$

IV. Evaluation of Application to Humanoid Robot

This section presents the evaluation of our voice-manipulation system with a humanoid robot, HRP-2 [11]. The evaluation was carried out in a robot-to-robot dialogue situation. Experimental results show how much information on the directionality in speech signals is delivered from HRP-2 to another humanoid robot, Robovie-R2, at various distances. We used Robovie-R2 for the evaluation because a state-of-the-art robot audition system HARK is implemented on Robovie-R2. Experiments such as simultaneous speech recognition have been conducted using Robovie-R2 [12].

A. Experiment setup

HRP-2 had a pair of stereo loudspeakers located at its waist as shown in Figure 10. The space between the speakers was 50 (cm). HRP-2 and Robovie-R2 stood face-to-face with a distance d in a room.The experiments were carried out with $d = 50, 100, 150$ (cm) which respectively correspond to intimate, personal, and social distances according to the Proxemics [13]. Speech signals are generated by a speech synthesizer developed by Fujitsu Ltd. The sentence used for this experiment is an excerpt from phonetically balanced sentences in Japanese.

B. Yaw-axis directionality

Robovie-R2 detected the direction from which the voice signal of HRP-2 was cast using a MUSIC algorithm implemented in a robot audition system called HARK [12]. By this algorithm, Robovie-R2 is able to detect the sound source direction from Robovie-R2's view with its spacial resolution 5°. Robovie-R2 has eight microphones around its head for sound localization as shown in Figure 11.

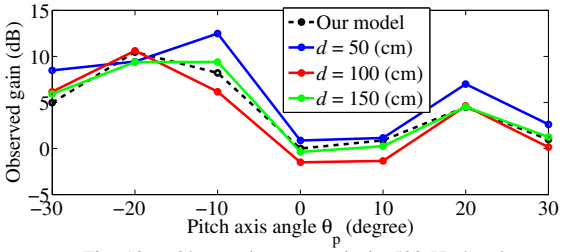

Fig. 13. Observed power ratio in 500 Hz band

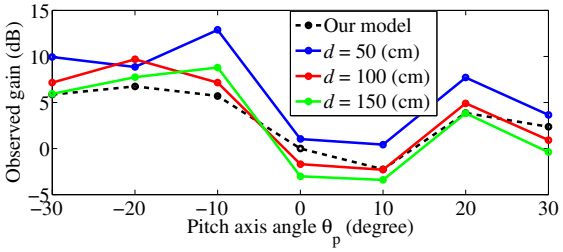

Fig. 14. Observed power ratio in 800–1000 Hz band

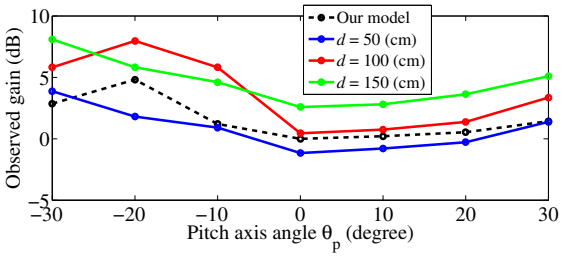

Fig. 15. Observed power ratio in 2500–4000 Hz band

Figure 12 shows the results for three distances d. Negative localization angle means left from Robovie-R2's view, meaning HRP-2 is was facing rightward. When the two robots were 100 or 150 (cm) away, Robovie-R2 perceived only 5° or 10° difference because the loudspeakers were placed as close as 50 (cm) from each other. However, Robovie-R2 successfully recognized which side HRP-2 was facing as long as HRP-2 rotated its head by more than 40° even when Robovie-R2 was 150 (cm) away. According to Figure 7, $g^y_{L,50^\circ} - g^y_{R,50^\circ} \approx 6$ (dB) is necessary to let Robovie-R2 perceive the directionality when it is 150 (cm) away.

We can conclude the gain derived from the inverse-square law is insufficient to present the directionality because the maximum difference in left-right channel gains was less than 5 (dB). Furthermore, it should be noted that the MUSIC algorithm detects only one sound source instead of detecting two sources independently from both speakers. Thus, we confirm that the use of stereo speakers is an appropriate way to present the azimuth directionality.

C. Pitch-axis directionality

For the evaluation of pitch-axis directionality, speech signals from HRP-2 were recorded with the front microphone attached to Robovie-R2. Recorded signals were put through

the three band-pass filters as written in equation (2). The plain speech signals generated by the speech synthesizer were also band-passed. Figures 13–15 show the difference of sound-pressure level between the recorded speech signals and the plain speech signals. Figures 13–15 indicate 500 Hz, 800–1000 Hz, and 2500–4000 Hz band, respectively. The black dotted plots indicate the gain of our model shown in Figure 4. When the shape of a gain curve is flat, it means the effect of the manipulation is less distinct; when the shape is close to that of black plots, the manipulation is well conveyed. The pitch angle θ_p ranging from -30° to 30° is the range of motion for HRP-2.

The shape of the gain curves in Figure 13 and 14 roughly conforms to our model shown as black plots. The gain for 50 (cm) distance is larger than those for 100 (cm) and 150 (cm) because Robovie-R2 is able to observe a louder signal when it stands near from HRP-2.

By contrast, the shape of the gain curve in Figure 15 is flat and different from our model. Namely, the observed voice manipulation was less distinct. The reason are twofold:

1) the sound-pressure level in a speech signal in 2500–4000 Hz region is relatively low and
2) the effect of white noise from A/D converter or stationary noise such as a cooling fan becomes dominant.

V. Conclusion

This paper presented a voice manipulation method consistent with a robot's head movements and posture. We assume that two kinds of head rotation affect the voice quality independently. That is, the yaw-axis rotation corresponds to the direction a voice is cast on the azimuth plane, while the pitch-axis rotation modulates the spectral envelope of speech signals due to the changes in a vocal tract. A voice is cast in a specific direction by using a pair of stereo speakers. The left-right sound-pressure balance is modeled by measuring impulse responses with a pair of microphones. We obtain the spectral envelope model for pitch-axis head movements on the basis of analysis of actual human vocalizations.

Our voice-manipulation architecture is capable of coping with unexpected body movements that may make the voice inconsistent with the posture. The switching of voice signals is a necessary function for speaking robots because they may encounter barriers that hinder them from moving as planned.

The experimental results prove that our method presents striking changes in voice quality and directionality in a robot-to-robot dialogue situation when robots stand an intimate distance (50 cm) from each other. We also confirm that an observable directionality declines as the distance between two robots becomes larger (150 cm).

Future works are as follows. First, adding vocal emission characteristics to our model is a promising way to improve the presentation of the directionality. On top of the spectral-envelope modulation, the effects of a transfer function between the talker and the listener will be presented. One approach to model the vocal emission characteristics is to use a microphone array surrounding a subject and to record his speech signals. A vertical microphone array will provide a model that will enable a further modulation corresponding to pitch-axis head motions other than spectral-envelope modification. Second, subjective evaluations should be carried out as to how uncomfortable people feel if there is an inconsistency between visual information and auditory information. The findings obtained by this research contribute to the knowledge of how much we should elaborate on voice-posture relationships.

Another future work is top-down modeling of the relationship between vocal tract and physical movements or between vocal band, the source in a source-filter model, and postures. This includes the verification of our most important assumption that pitch and yaw motions affect the voice independently. As far as the authors know, psychoacoustic observation about motion-speech relationship has not been specified. Applicable approaches are using X-ray imaging or magnetic resonance imaging to visualize vocal organs even while a subject is speaking.

ACKNOWLEDGMENTS This research was partially supported by MEXT, Grant-in-Aid for Scientific Research.

References

[1] S. Fujie, D. Watanabe, Y. Ichikawa, H. Taniyama, K. Hosoya, Y. Matsuyama, and T. Kobayashi. Multi-modal integration for personalized conversation: Towards a humanoid in daily life. In *8th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2008)*, pages 617–622, Dec. 2008.

[2] D. Matsui, T. Minato, K. F. MacDorman, and H. Ishiguro. Generating natural motion in an android by mapping human motion. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3301–3308, Aug. 2005.

[3] R. Dillmann, R. Becher, and P. Steinhaus. Armar II - a learning and cooperative multimodal humanoid robot system. *International Journal of Humanoid Robotics*, 1(1):143–155, 2004.

[4] Z. Inanoglu and S. Young. Intonation modelling and adaptation for emotional prosody generation. In *Affective Computing and Intelligent Interaction*, pages 286–293, 2005.

[5] D. Erickson. Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology*, 26(4):317–325, 2005.

[6] T. Tasaki, S. Matsumoto, H. Ohba, M. Toda, K. Komatani, T. Ogata, and H. G. Okuno. Distance based dynamic interaction of humanoid robot with multiple people. *Innovations in Applied Artificial Intelligence, LNAI*, 3533:111–120, 2005.

[7] Kenichi Aoki, T. Kamakura, and Y. Kumamoto. Parametric loudspeaker – characteristics of acoustic field and suitable modulation of carrier ultrasound. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 74(9):76–82, 2007.

[8] P. Birkholz, D. Jackèl, and B. J. Kröger. Construction and control of a three-dimensional vocal tract model. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, pages 873–876, 2006.

[9] H. Kawahara, M. Morise, R. Nisimura, T. Irino, and H. Banno. Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP '08)*, pages 3933–3936, 2008.

[10] G. Fant. *Acoustical Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. The Hague, Mouton, 1970.

[11] K. Kaneko, F. Kanehiro, S. Kajita, H. Hirukawa, T. Kawasaki, M. Hirata, K. Akachi, and T. Isozumi. Humanoid robot HRP-2. In *IEEE International Conference on Robotics and Automation*, volume 2, pages 1083–1090 Vol.2, 26-May 1, 2004.

[12] K. Nakadai, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. An open source software system for robot audition hark and its evaluation. In *Humanoids 2008*, pages 561–566, Dec. 2008.

[13] E. T. Hall. *Hidden Dimension*. Doubleday Publishing, 1996.