PALADYN Journal of Behavioral Robotics

VERSITA

Research Article · DOI: 10.2478/s13230-010-0009-x · JBR · 1(1) · 2010 · 80-88

Voice-awareness control for a humanoid robot consistent with its body posture and movements

Takuma Otsuka[1]*,
Kazuhiro Nakadai[2 3]†,
Toru Takahashi[1],
Kazunori Komatani[1],
Tetsuya Ogata[1],
Hiroshi G. Okuno[1]

1 Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan

2 Honda Research Institute Japan, Co., Ltd., Wako, Saitama, 351-0114, Japan,

3 Graduate School of Information Science and Engineering, Tokyo Institute of Technology

Received 21 February 2010

Accepted 19 March 2010

Abstract

This paper presents voice-awareness control consistent with robot's head movements. For a natural spoken communication between robots and humans, robots must behave and speak the way humans expect them to. The consistency between the robot's voice quality and its body motion is one of the most especially striking factors in naturalness of robot speech. Our control is based on a new model of spectral envelope modification for vertical head motion, and left-right balance modulation for horizontal head motion. We assume that a pitch-axis rotation, or a vertical head motion, and a yaw-axis rotation, or a horizontal head motion, effect the voice quality independently. The spectral envelope modification model is constructed based on the analysis of human vocalizations. The left-right balance model is established by measuring impulse responses using a pair of microphones. Experimental results show that the voice-awareness is perceivable in a robot-to-robot dialogue when the robots stand up to 150 cm away. The dynamic change in the voice quality is also confirmed in the experiment.

Keywords

voice awareness · robot speech signal control · 2D voice manipulation · source filter model · human robot interaction

1. Introduction

We have an increasing number of chances to have verbal and physical interactions with robots thanks to the development of robots intended to interact with humans. For example, ROBISUKE [9] interacts with multiple persons by voiced speech and physical behaviors; Repliee Q2 [20] has a vivid human-like appearance to help people become more familiar with it and its behavior; and ARMAR II [6] performs human-like operations in household tasks. They can speak very fluently and naturally thanks to up-to-date text-to-speech (TTS) systems. A TTS engine synthesizes wave files of voiced speech given a text, and a number of TTS engines are available, either as commercial or free software, e.g., Microsoft, IBM, or CMU's Festival system [4]. One of the hot topics concerning TTS is how to improve the quality of synthesized speech with respect to emotion [7]. This is because natural and successful verbal interactions between humans and robots depend highly on the quality of synthesized speech. In addition to speech quality, robots must behave and speak the way humans expect them to. For example, robots should face the talker or give back-channel feedback with proper timing [3].

The consistency between the robot's voice quality and its body posture or motion is one of the most especially striking factors in the naturalness of robot speech [23]. When the robot faces upward, the voice should sound strong and clear; when the robot bends down, the voice should become weak and vague. We refer to this consistency as ***voice-awareness*** [19]. Here, voice-awareness is defined as a change in the voice influenced by body movements or posture. The voice-awareness helps us be aware of the physical disposition of the robot. We can infer the direction in which the robot is talking or whether the robot is moving its body or not by hearing the robot's voice.

Changes in the voice quality corresponding to physical posture or motions are classified as a 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 the speaker's states, both internal and external ones; the former includes the speaker's feelings and the latter includes a speaker's physical posture.

Existing studies intended to add paralinguistic information to speech signals focus on physically-independent features such as intonation [11] or emotional aspects [7, 16]. These studies provide spoken dialogue systems with natural speech sounds, and as a result, we find it comfortable to use such systems. To apply them to robots, these kinds of paralinguistic information should incorporate mechanisms for coordinating with robot's movements or posture, because the changes in voice quality caused by their body movement and posture is ignored.

An ultradirectional ultrasound loudspeaker may facilitate voice-awareness control, because it generates a narrow beam of speech voices. The humanoid, SIG2, is equipped with a "champion-belt" like ultradirectional loudspeaker at its waist [25]. When it turns, its voiced speech follows the body movements in the azimuth plane. Since the

*E-mail: ohtsuka, tall, komatani, ogata, okuno@kuis.kyoto-u.ac.jp
†E-mail: nakadai@jp.honda-ri.com

direction of speech sounds is casted on the azimuth plane, the consistency between the direction of speech sounds and the robot's face motion is restricted to the azimuth movements of the robot. The humanoid, Honda ASIMO, is equipped with a small ultradirectional loudspeaker at the position of its mouth [21]. ASIMO can generate a sound beam of about 20 degrees to deliver voice-awareness. In other words, a humanoid achieves the consistency between the direction of speech sounds and the robot's face motion. However, critical problems remain. Firstly, the sound from them has little power in frequency bands less than 500 Hz due to their mechanism [1]. Secondly, this method only addresses 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.

Another solution to the quality control of speech sounds may use a 3D sound authoring tool. Kim *et al.* use the SoundLocus tool [5] to evaluate their robot audition system installed at SIG2 [18]. The problem is that such authoring tools are dedicated to a particular listener, not an audience around the robot. Another problem is a lack of real-time processing capability.

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. This method manipulates the voice signal in accordance with the vertical head motion and the horizontal head motion under the assumption that these two motions affect the voice independently. The model for the vertical head motions is constructed through the one-third octave band analysis of human speech sounds. The rest of the paper is organized as follows: Section 2 presents a voice-awareness architecture to keep robot's voice consistent with its body posture. Section 3 presents the measurements of parameters for voice quality manipulation based on the source-filter theory of speech synthesis. Section 4 describes the algorithm of voice quality manipulation. Section 5 evaluates the voice-awareness control by using two robots. Finally Section 6 concludes the paper with mentioning future works.

2. Voice-Awareness Architecture to keep robot's voice consistent with its body posture

This section presents a voice-awareness architecture to keep the robot's voice consistent with its body posture. Then the voice-awareness problem for humanoid robots is clarified.

2.1. Voice-Awareness Architecture

Figure 1 outlines our envisioned architecture capable of controlling voice-awareness and coping with unexpected body movements that can cause the voice to be inconsistent with the posture. The voice-awareness control 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 that generates a plain voice signal without voice-awareness. The motion command is delivered to the parts of the robot body for their movements. The voice-awareness is attached to the voice signal based on the actual body posture such as joint angles acquired by sensors attached to the joints.

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 abrupt change in its posture can make the voice inconsistent with the robot's actual posture. An exceptional exclamation such as "Ouch !" or groan should be produced in such incidents to inform surrounding people of troubles. The exceptional exclamation database in Figure 1 stores such specific voice signals. When the monitor that compares the actual posture data and the planned motion detects an unexpected body movement, the robot speech switches to a specific phrase.

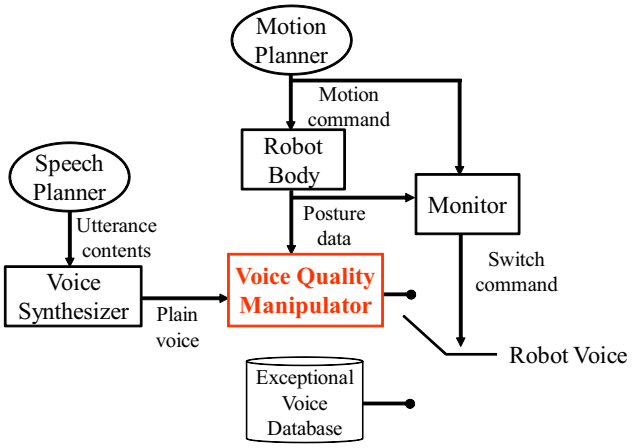

Figure 1. *Voice-awareness architecture*

2.2. 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 [2]. However, this method has just generated clear consonants, and, it still has difficulty in generating long words where sophisticated control of vocal tract parameters is required. In addition, the manipulation should be feasible and applicable to any words at a low computational cost to manipulate the voice signals in real-time as joint angles are monitored. STRAIGHT [15] 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 2 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. This assumption of the independency is made to facilitate the voice manipulation process. The pitch rotation changes the spectral envelope of the speech signal because this movement al-

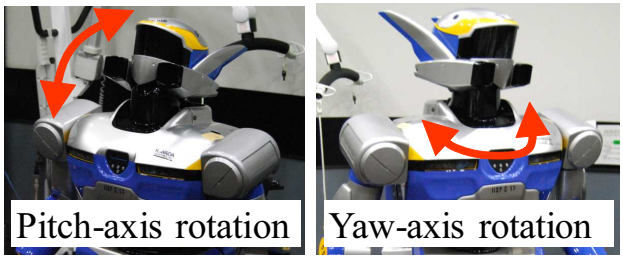

Figure 2. Head motions posed by HRP-2. Pitch-axis on the left and Yaw-axis on the right.

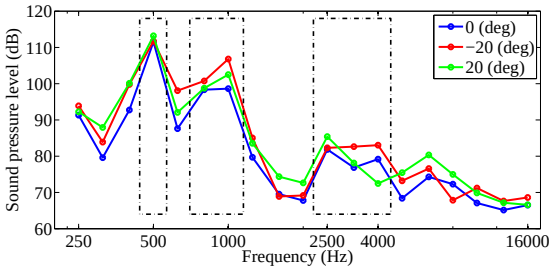

Figure 4. Power at each frequency band with 0-degree vocalization

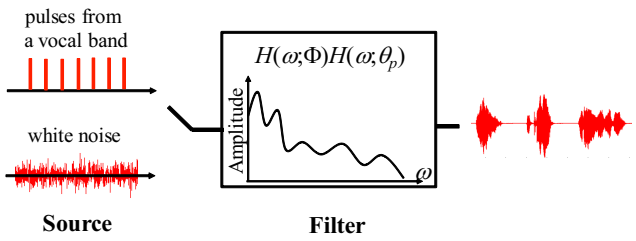

Figure 3. Source filter model of human voice

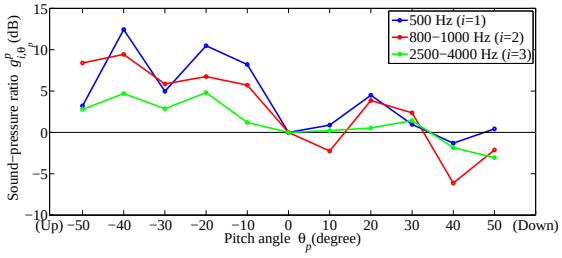

Figure 5. θ_p–gain model for three bands

ters the vocal tract, which works as an acoustic filter in accordance with the source filter model [8]. The bending of the vocal tract by the pitch axis rotation rather than the twisting by the yaw axis rotation is considered dominant in the spectral characteristics of the voice signal. 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-position consistent speech signal $\hat{x}(t)$,
Assumption: θ_p and θ_y affect $x(t)$ independently,

where t signifies 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.

3. Measurement of parameters for voice manipulation

Our method manipulates the spectral envelope of the voice signal in accordance with pitch-axis angle θ_p and modulates the balance of sound pressure level of left-right channel in accordance with yaw-axis angle θ_y. This section presents the parameters for these manipulations based on the measurement of human vocalizations for the pitch axis and impulse response of two microphones for the yaw axis.

3.1. Source Filter Model

We assume that the pitch rotation θ_p effects the filter in the source filter model. Figure 3 outlines the source filter model of a human voice [8]. 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 effect on the voice modulated by the shape of the vocal tract. The frequency response of the filter can be divided into two components: $H(\omega;\Phi)$ and $H(\omega;\theta_p)$, where ω indicates a frequency and Φ is a type of phoneme. $H(\omega;\Phi)$ shapes formants or other phonemic features. $H(\omega;\theta_p)$ is the effect of vertical head motions.

Note that these filters are time-variant because Φ and θ_p change depending on the time. θ_p is time-variant when the robot talks while it is moving its head. Our method first constructs a model of $H(\omega;\theta_p)$ for static θ_p and applies the model to voice signals for varying θ_p.

3.2. The construction of $H(\omega;\theta_p)$

We build a spectral envelope model of $H(\omega;\theta_p)$ by inspecting a human voice for various angles θ_p. The recording set up is as follows. Speech signals of a male subject, one of the authors, were recorded with a close-talking microphone in an anechoic chamber as shown in Figure 6. A 10-second-long sweep-tone vocalization of the vowel /a/ was recorded with the subject's head moving 10 degrees at a time from 50 degrees downward to 50 degrees upward. A sweep tone was used to avoid the effect of the fundamental frequency and ranged from 261 Hz to 523 Hz, which corresponds to the musical note C. The recorded

voice signal was then analyzed with one-third octave bands:

$$P(o,\theta_p) = \int |\bar{U}(\omega,\theta_p)|^2 Q(\omega,o)\mathrm{d}\omega, \quad (1)$$

$$Q(\omega,o) = \begin{cases} \frac{\omega-f_{o-1}}{f_o-f_{o-1}} & (f_{o-1} \le \omega < f_o), \\ \frac{f_{o+1}-\omega}{f_{o+1}-f_o} & (f_o \le \omega < f_{o+1}), \\ 0 & \text{otherwise}, \end{cases} \quad (2)$$

where $P(o,\theta_p)$ denotes the power of octave band index o with the pitch axis being θ_p degree, and $\bar{U}(\omega,\theta_p)$ denotes normalized amplitude for frequency ω the pitch axis θ_p. To emphasize changes in the spectral envelope without changes in the power, the recorded signal $U(\omega,\theta_p)$ is normalized to $\bar{U}(\omega,\theta_p)$ such that $\int \bar{U}(\omega,\theta_p)^2\mathrm{d}\omega$ becomes constant for various θ_p. $Q(\omega,o)$ denotes the window function in the frequency domain for octave band o. f_o is the center frequency of o-th octave band and defined as $500 \times \sqrt[3]{2}^{o-3}$ Hz. The octave band index o ranged from 0 to 18. Sound pressure levels for each band compared to the respective levels at 0 degrees, $S(o,\theta_p)$, were calculated.

$$S(o,\theta_p) = 10\log_{10}\frac{P(o,\theta_p)}{P(o,0)}. \quad (3)$$

Figure 4 shows the results of one-third octave band analysis of the voice signals where, $\theta_p = 0, -20, 20$ degrees. Negative pitch angles indicate facing upward whereas positive ones indicate facing downward. We choose three frequency bands marked with black dashed lines (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 because most formants exist in these frequency region.
2. The equal-loudness contours [13] reveal that the auditory perception of humans is sensitive to changes in these frequency regions.
3. More change in the sound pressure level is observed varying θ_p in these bands. The formants in the second and the third frequency bands are reported as striking factor in singing voice [26].

Therefore, these bands are considered most effective for the voice-quality manipulation. Figure 5 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 gains g^p_{i,θ_p} are calculated as follows:

$$g^p_{i,\theta_p} = 10\log_{10}\frac{C_i(\theta_p)}{C_i(0)}, \quad (4)$$

$$C_i(\theta_p) = \int |\bar{U}(\omega,\theta_p)|^2 W_i(\omega)\mathrm{d}\omega, \quad (5)$$

$$W_i(\omega) = \begin{cases} \frac{\omega-b^0_i}{c^0_i-b^0_i} & (b^0_i \le \omega < c^0_i), \\ 1 & (c^0_i \le \omega < c^1_i), \\ \frac{b^1_i-\omega}{b^1_i-c^1_i} & (c^1_i \le \omega < b^1_i), \\ 0 & \text{otherwise}, \end{cases} \quad (6)$$

where $C_i(\theta_p)$ denotes the energy for the i-th band with the pitch angle θ_p. The upper suffix p of the gain means that this is a gain for pitch angle, and i represents the band index. b^0_i and b^1_i are the bottom edges of the i-th frequency band and c^0_i and c^1_i are cutoff frequencies. These values are specified in Table 1.

Table 1. Edges of frequency bands

Denotations	Frequency Hz	
Bottom edges for $i=1$ (b^0_1, b^1_1)	400	625
Cutoff freq. for $i=1$ (c^0_1, c^1_1)	500	500
Bottom edges for $i=2$ (b^0_2, b^1_2)	625	1250
Cutoff freq. for $i=2$ (c^0_2, c^1_2)	800	1000
Bottom edges for $i=3$ (b^0_3, b^1_3)	2000	5000
Cutoff freq. for $i=3$ (c^0_3, c^1_3)	2500	4000

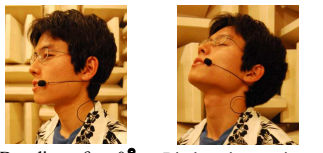

Figure 6. Recording of human vocalization

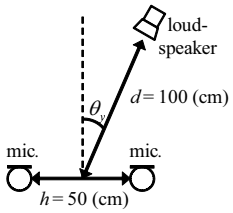

Figure 7. Setup for a left-right balance measurement

Observations of vocalizations by a 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 5.

3.3. Azimuth plane control

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 limits the sound quality and volume from the loudspeaker.

We confirmed the left-right balance by measuring impulse responses in an anechoic chamber as shown in Figure 7. The angle θ_y ranged from $-90°$ to $90°$, every $10°$, where $0°$ is the center position and a positive angle means the left direction. 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 8. Signals from the two speakers are assumed to be planar waves for the derivation of the plot for the inverse-square law. The y axis represents the sound-pressure ratio compared to the 0-degree sound level. Both plots indicate the left channel. We use the measured result rather than the simulated result because an exaggerated modification is necessary to show the directional information clearly. The difference between the two curves shown in Figure 8 is produced because the assumption of planar wave

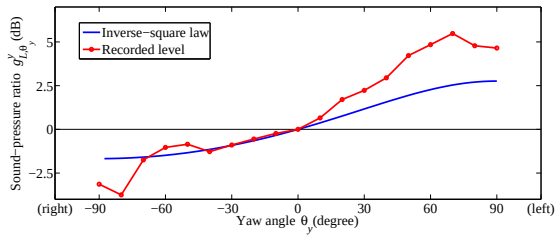

Figure 8. Empirical and theoretical θ_y–gain model for the left channel

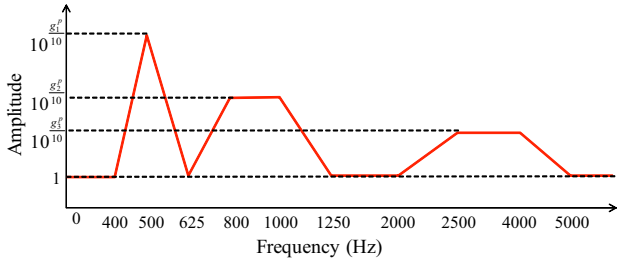

Figure 9. The amplitude of spectral envelope for pitch axis $H^p(\omega;\theta_p(\tau))$

scarcely holds in the case a sound source is only 100 cm away from the microphones.

4. 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 monaural 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.

4.1. Pitch-axis modification

The spectral envelope modification is carried out in time-frequency domain. The input signal $x(t)$ is converted into corresponding complex representation in time-frequency domain $X(\omega,\tau)$ by short-time Fourier transform (STFT) where τ denotes a discrete time. The step size of τ is determined by the step size of STFT. The Hamming window is employed as a window function. The window size is L ms and the step sizes is S ms.

At time τ, the pitch angle at the time $\theta_p(\tau)$ is acquired for the spectral envelope modification. The gains for three frequency bands are interpolated on a dB scale in Figure 5.

$$g_i^p(\tau) = \frac{g^p_{i,\theta_m}(\theta_{m+1}-\theta_p(\tau)) + g^p_{i,\theta_{m+1}}(\theta_p(\tau)-\theta_m)}{10}, \quad (7)$$

$$\theta_{m+1} = (\lfloor \theta_p(\tau)/10 \rfloor + 1)\times 10, \quad (8)$$

$$\theta_m = (\lfloor \theta_p(\tau)/10 \rfloor)\times 10, \quad (9)$$

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 floor function. For example, when $\theta_p(t) = 35°$, $\theta_{m+1} = 40°$ and $\theta_m = 30°$, consequently, $g_i^p(t) = (g^p_{i,30°} + g^p_{i,40°})/2$.

Then, a spectral envelope modification filter $H^p(\omega;\theta_p(\tau))$ is built using three gains g_i^p, $i = 1,2,3$ as Eq. (10). Figure 9 shows the amplitude of $H^p(\omega;\theta_p(\tau))$. This filter has three peaks in 500, 800–1000, 2500–4000 Hz regions.

$$H^p(\omega;\theta_p(\tau)) = 1 + \sum_{i=1}^{3} H_i^p(\omega;g_i^p(\tau)), \quad (10)$$

$$H_i^p(\omega;g_i^p) = \begin{cases} (10^{g_i^p}-1)\frac{\omega - b_i^0}{c_i^0 - b_i^0} & (b_i^0 \le \omega < c_i^0), \\ 10^{g_i^p} - 1 & (c_i^0 \le \omega < c_i^1), \\ (10^{g_i^p}-1)\frac{b_i^1 - \omega}{b_i^1 - c_i^1} & (c_i^1 \le \omega < b_i^1), \\ 0 & \text{otherwise}, \end{cases} \quad (11)$$

The frequencies b_i^0, b_i^1, c_i^0, and c_i^1 are specified in Table 1. These frequencies are center frequencies for one-third octave band analysis. After constructing $H^p(\omega;\theta_p(\tau))$, the input signal $X(\omega,\tau)$ is amplified as

$$X^p(\omega,\tau) = X(\omega,\tau)H^p(\omega;\theta_p(\tau)). \quad (12)$$

4.2. Yaw-axis modification

The pitch-axis modulated and monaural signal $X^p(\omega,\tau)$ is first doubled to a stereo signal $\mathbf{X}(\omega,\tau) = [X_L^p(\omega,\tau)\ X_R^p f\omega,\tau)]$, both channels of which equal $X^p(\omega,\tau)$. Both channels of the stereo signal, $x_L(t)$ and $x_R(t)$, are amplified in accordance with the control model shown in Figure 8.

The yaw angle at time τ is obtained as $\theta_y(\tau)$. Then, the gain for each channel $g_j^y(\theta_y(\tau))$ $(j = L,R)$ is calculated by interpolating the discretely measured gains $g^y_{\theta_n}$ for continuous angle $\theta_y(\tau)$ as in equations (13) and (14).

$$g_L^y(\tau) = \frac{g^y_{\theta_n}(\theta_{n+1}-\theta_y(\tau)) + g^y_{\theta_{n+1}}(\theta_y(\tau)-\theta_n)}{10}, \quad (13)$$

$$g_R^y(\tau) = g_L^y(t;-\theta_y(\tau)), \quad (14)$$

$$\theta_{n+1} = (\lfloor \theta_y(\tau)/10 \rfloor + 1)\times 10, \quad (15)$$

$$\theta_n = (\lfloor \theta_y(\tau)/10 \rfloor)\times 10. \quad (16)$$

The gains of the left and right channels are symmetric as expressed in equation (14). In the next step, each channel is amplified by the respective gain as

$$X_j^y(\omega,\tau) = X_j^p(\omega,\tau)\times 10^{\frac{g_j^y(\tau)}{10}} (j = L,R). \quad (17)$$

Finally, the stereo complex signal $\mathbf{X}^y(\omega,\tau) = [X_L^y(\omega,\tau)\ X_R^y(\omega,\tau)]$ is converted into a time domain signal using the inverse Fourier transform.

4.3. Real time implementation

Our algorithm is able to run in real time when the input signal is captured on-line by a microphone or other devices, however some inevitable latency exists. The latency consists of (1) buffering of the input signal to be processed and (2) processing time itself. The processes from Eq. (7) to Eq. (17) are carried out every S ms, where S is the step size for STFT. For the incremental process, the system waits for the input signal to be buffered for S ms. To produce an output signal in real time, these processes should be carried out within S ms. Therefore, the delay in the output signal compared to the input is between S and $2S$ ms. In our implementation, STFT parameters L and S are set as: $L = 32$ and $S = 16$ ms, where the sampling rate is 16 kHz. The reasons are twofold: (1) the preferred frame length for speech signal analysis is around 30 ms, because speech signals are regarded as stationary within this length, and (2) the estimated latency between 16 and 32 ms is acceptable for the use of verbal dialogue.

Our system is implemented on Windows Vista with Core2Duo, 1.6 GHz CPU. The processing time for a 32-millisecond frame is approximately 4 ms. Therefore, the latency of the voice manipulation is around 20 ms. If a robot has to vocalize in synchronization with other voice, such as singing a chorus, this latency should be avoided and prediction of posture may be one possible solution.

5. Evaluation of application to humanoid robots

This section presents the evaluation of our voice-awareness control with two types of humanoid robots: HRP-2 [14] and HIRO [12]. The validity of our method should be confirmed from the physical and psychological viewpoints. As a first step of such evaluations, we focus on the physical aspects. In other words, we use a robot head with 8 microphones to ensure a precise localization of generated sounds, although the robot head introduces another problems caused by a transfer functions different from humans'. Needless to say, head related transfer function depends on each person's head shape, hair style and cloths. We believe that this evaluation method provides rough characteristics of our method. The psychological evaluation of our method is one of important future works. 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 or HIRO to another humanoid robot, Robovie-R2, at various distances. In the following experiments, HRP-2 or HIRO is the speaker and Robovie-R2 is the listener.

Our voice-awareness control is evaluated from three aspects:

1. the yaw-axis directionality validation using a sound localization algorithm,
2. the static spectral envelope manipulation regarding the pitch axis,
3. the dynamic spectral envelope manipulation.

HRP-2 is used for the first two experiments to evaluate the static feature of the yaw-axis stereo gain control and pitch-axis spectral envelope manipulation. HRP-2 and HIRO are used for the third experiment. The difference in the range of capable joint movement between two robots is observed in the manipulated voice signals.

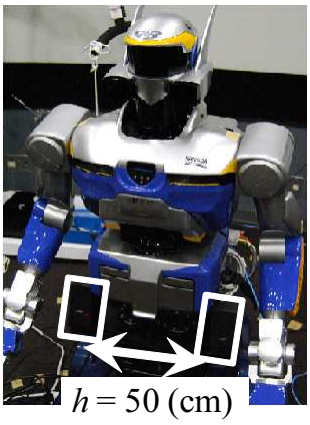

Figure 10. HRP-2 with its stereo loudspeakers marked by rectangles

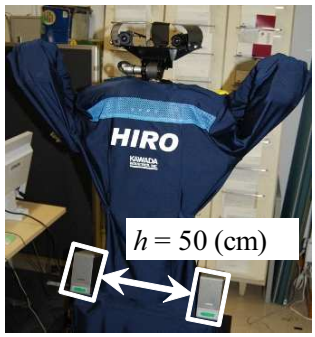

Figure 11. HIRO with its stereo loudspeakers

Figure 12. HRP-2 on the left and Robovie-R2 on the right

5.1. Experimental 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 separated by a distance d in a room. The speech signals to manipulate were excerpts from JNAS phonetically balanced sentences in Japanese. The first and the second experiments were carried out with $d = 50, 100, 150$ cm which respectively correspond to intimate, personal, and social distances according to the Proxemics [10].

The third experiment was carried out with $d = 100$ cm. Figure 13 shows the sequence of pitch-axis motion made by HRP-2 and HIRO. The range of pitch-axis motion of HRP-2 is between -30 and 30 deg whereas that of HIRO is limited to angles above 0 deg. In other words, HIRO is unable to face upward.

5.2. Experiment 1: 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 [22]. Using this algorithm, Robovie-R2 is able to detect the sound source direction from Robovie-R2's view with its

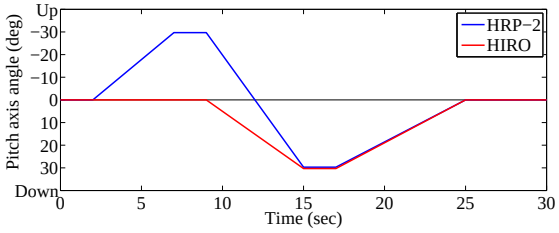

Figure 13. Pitch-axis motion trajectory for two robots

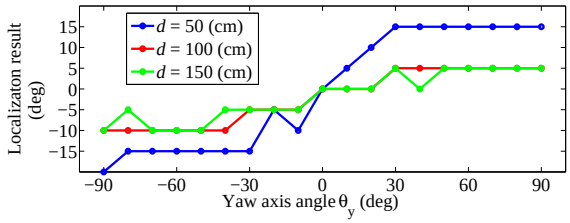

Figure 14. Sound localization result on the azimuth plane

spacial resolution of 5°. Robovie-R2 has eight microphones around its head for sound localization as shown in Figure 12.

Figure 14 shows the results for three distances d. Negative localization angles signify left from Robovie-R2's view, or that HRP-2 is rightward facing. When the two robots were 100 or 150 cm apart, Robovie-R2 only perceived 5° or 10° differences because the loudspeakers were placed as close as 50 cm to 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 8, $g^y_{L,50^\circ} - g^y_{R,50^\circ} \approx 6$ dB is necessary for Robovie-R2 to perceive directionality when it is 150 cm away.

5.3. Experiment 2: Static 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 whose passbands are 500 Hz, 800–1000 Hz, 2500–4000 Hz, respectively. The pitch angle θ_p ranged from -30° to 30° by 10 deg. For each θ_p, the energy that each band-pass filter outputs is compared with the energy from the corresponding band-pass filter with $\theta_p = 0$ deg. Figures 15–17 show the observed gain in the sound-pressure level compared with the recorded signal when $\theta_p = 0$ deg. Figures 15–17 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 5. When an observed gain is close to 0 dB, it means the effect of the manipulation is less distinct; when the gain is close to black plots, the manipulation is well conveyed.

These figures show the effect of voice manipulation is generally delivered when the listener robot stands 150 cm away. The manipulation is less transfered when $\theta_p = -20$ deg. This may be because the signal is so amplified that the micorphone attached to robovie failes to catch the dyanmics of the voice signal amply.

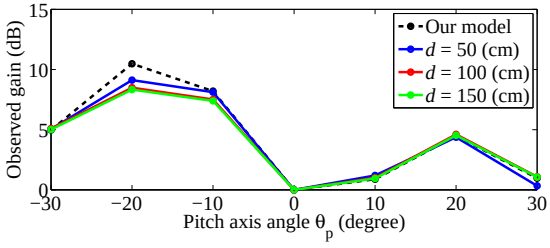

Figure 15. Observed power ratio in 500 Hz band

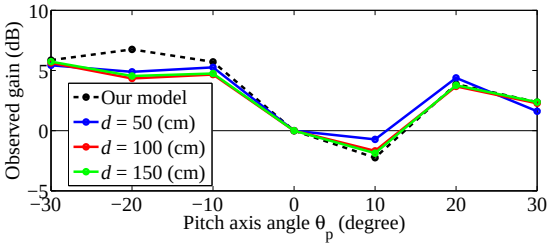

Figure 16. Observed power ratio in 800–1000 Hz band

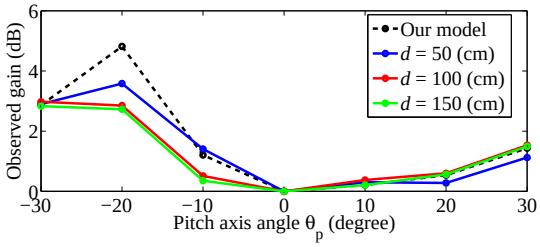

Figure 17. Observed power ratio in 2500–4000 Hz band

5.4. Experiment 3: Dynamic pitch-axis directionality

Two robots, HRP-2 and HIRO, uttered phonetically balanced sentences in Japanese while moving their heads as shown in Figure 13. The trajectories of the two robots are different because of the limited range of motion of HIRO. The voice is manipulated in accordance with the actual robot pitch-axis angle.

The analysis is carried out as follows. The voice from two robots are recorded with a microphone 100 cm away from the speaker robot with the sampling rate 16 kHz. The voice without the manipulation is also recorded. These recordings are converted into time-frequency domain by STFT with the 32 ms long Hamming window and 16 ms long step. For each time frame, the energy for the three bands $E^i_r\ \tau$ are calculated as follows:

$$E^r_i(\tau) = \int |X^r_{obs}(\omega,\tau)|^2 W_i(\omega)\mathrm{d}\omega, \qquad (18)$$

where $|X^r_{obs}(\omega,\tau)|^2$ is the power at ω Hz and time frame τ of the observed voice signal for robot r. r denotes the index of the robot; HRP-2 or HIRO. W_i represents the band-pass window in Equation 6. The edge frequencies, b^0_i, b^1_i, c^0_i, c^1_i, are specified in Table 1. The gain of the manipulated voice signal to the original voice signal is calculated for each

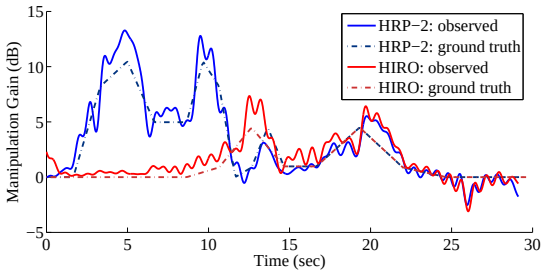

Figure 18. Observed power ratio in 500 Hz band

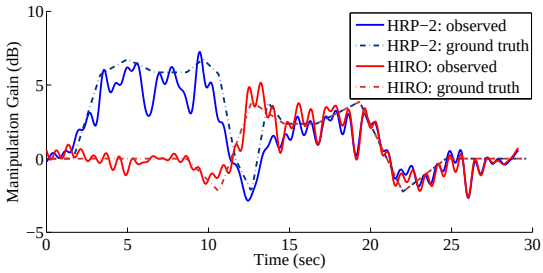

Figure 19. Observed power ratio in 800–1000 Hz band

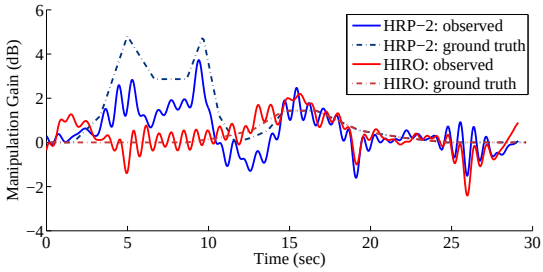

Figure 20. Observed power ratio in 2500–4000 Hz band

frame. The obtained gain curves contain huge fluctuations because the recorded signals contain a noise caused by caused by the connection of microphones or in the environment. Since the rough shape of these curves represents how the original voice signal is amplified with regard to its pitch angle, the fluctuations due to a noise should be ignored. Therefore, the curves are put through low-pass filter to reduce the noise from the gain curves.

Figures 18–20 show the gain curves and target gain curves based on our model in Figure 5 for each robot. Solid lines represent the observed gain curves and broken lines represent the target gain curves. The observed gain curves in Figures 18 and 19 roughly match the target gain curves. These results confirm that the voice manipulation is physically delivered when the robots poses in dynamic body motions. On the contrary, the observed curves in Figure 20 are different from the target curves. This is because the noise in the observed voice signal becomes dominant when the voice is acquired by the robot's microphones.

6. Conclusion and Future works

This paper presented a voice manipulation method consistent with a robot's head movements and posture to improve voice-awareness. 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 changes in the voice quality for the pitch-axis angle and the azimuth directionality for the yaw-axis angle in a robot-to-robot dialogue situation. Both the pitch-axis manipulation and the yaw-axis manipulation are physically observable when the distance between the speaker robot and the listener robot is up to 150 cm. The dynamic manipulation of the voice quality for moving pitch-axis angle is also confirmed in the experiment.

Future works are as follows. Firstly, 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.

Secondly, our method contributes to a vivid singing voice by a robot [24]. The singing voice is generated by a singing voice generator called Vocaloid [17]. Our method enables the robot singer to emphasize its own voice when the robot faces upward as an opera singer vocalizes loudly by straightening his/her own body.

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, the psychoacoustic observation of the motion-speech relationship has not been specified. Applicable approaches are using X-ray imaging or magnetic resonance imaging to visualize vocal organs while a subject is speaking.

Acknowledgment

This work was supported in part by the Grant-in-Aid for Scientific Research (S) and in part by the Global COE program. The authors would like to thank Mario Suzuki and the members of Okuno and Ogata Laboratory for their discussion and valuable suggestions.

References

[1] K. 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.

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

[3] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99), pages 1146–1151, 1999.

[4] R. A. J. Clark, K. Richmond, and S. King. Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication, 49(4):317–330, 2007. .

[5] ARNIS Sound Technologies Co., Ltd. Soundlocus. http://www.arns.com/english/tech1.html, 2009.

[6] 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.

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

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

[9] 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. .

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

[11] Z. Inanoglu and S. Young. Intonation modelling and adaptation for emotional prosody generation. Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science 3784:286–293, 2005.

[12] Kawada Industries, Inc. Upper body humanoid robot HIRO. http://global.kawada.jp/mechatronics/hiro.html, 2009.

[13] ISO. ISO 226:2003: Acoustics – Normal equal-loudness-level contours. International Organization for Standardization, 2003.

[14] 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 (ICRA-2004), volume 2, pages 1083–1090 Vol.2, 26-May 1, 2004.

[15] 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 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08), pages 3933–3936, 2008.

[16] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno. Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09), pages 3377–3680, 2009.

[17] H. Kenmochi and H. Ohshita. Vocaloid – commercial singing synthesizer based on sample concatenation. In Proceedings of INTERSPEECH, pages 4010–4011, 2007.

[18] H. D. Kim. Binaural Active Audition for Humanoid Robots. PhD thesis, Graduate School of Informatics, Kyoto University, Sep. 2009.

[19] Y. Kubota, M. Yoshida, K. Komatani, T. Ogata, and H. G. Okuno. Design and implementation of 3D auditory scene visualizer towards auditory awareness with face tracking. In IEEE International Symposium on Multimedia (ISM2008), pages 468–476, Dec. 2008.

[20] 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 (IROS-2005), pages 3301–3308, Aug. 2005.

[21] K. Nakadai and H. Tsujino. Towards new human-humanoid communication: Listening during speaking by using ultrasonic directional speaker. In IEEE International Conference on Robots and Automation (ICRA-2005), pages 1483–1488, Apr. 2005.

[22] 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 8th IEEE-RAS International Conference on Humanoids (Humanoids 2008), pages 561–566, Dec. 2008.

[23] T. Otsuka, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno. Voice quality manipulation for humanoid robots consistent with their head movements. In 9th IEEE-RAS International Conference on Humanoids (Humanoids-2009), pages 405–410, Dec. 2009.

[24] T. Otsuka, K. Nakadai, Toru Takahashi, K. Komatani, T. Ogata, and H. G. Okuno. Incremental Polyphonic Audio to Score Alignment using Beat Tracking for Singer Robots. In Proceedings of IEEE/RSJ Int'l Conference on Intelligent Robots and Systems, pages 2289–2296, 2009.

[25] 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, Lecture Notes in Artificial Intelligence 3533:111–120, 2005.

[26] A. Vurma and J. Ross. Where Is a Singer's Voice if It Is Placed "Forward". Journal of Voice, 16(3):383–391, 2002.