Voice-Awareness Control Consistent with Robot's Body Movements

*Takuma Otsuka[†], Kazuhiro Nakadai[‡], Toru Takahashi[†],
Kazunori Komatani[†], Tetsuya Ogata[†], Hiroshi G. Okuno[†]
† Graduate School of Informatics, Kyoto University
‡ Honda Research Institute Japan Co., Ltd.

Abstract— This paper presents voice-awareness control related to robot's head movements. Our control is based on a new model of spectral envelope modification for the vertical head motions, and left-right balance modulation for the horizontal head motions. The spectral envelope modification model is 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 50 cm away. We also confirmed observable voice-awareness declines as the distance becomes large up to 150 cm.

Key Words: Voice awareness, Voice-quality manipulation, Robot's posture and movements

1. 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] or ARMAR II [2]. To realize natural and successful conversations between humans and robots, robots must behave and speak in a 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. This kind of consistency is referred to as *voice-awareness*. Here, voice-awareness is defined as a change in the voice corresponding to body movements and helps us be aware of the physical information of robots. This is a part of nonliteral information in speech signals. Voice-awareness control is essential to natural conversations. For instance, when the robot faces upward, the voice should sound strong and clear; when the robot bends down, the voice should become weak and vague.

Existing studies intended to add nonliteral information to speech signals focus on physically-independent features such as intonation [3] or emotional aspects [4]. These studies provide spoken dialogue systems with natural speech sounds, and as a result, we find it comfortable to use such systems. However, these kind of additional information is insufficient for robots because we are unaware of their body movement from their voice.

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 [5]. However, this approach encounters several problems such as deteriorated sound quality and a lack of change in the voice quality related to the robot's vertical head motion.

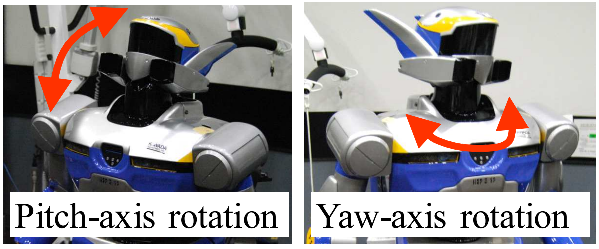

Fig.1: Head motions in question posed by HRP-2. Pitch-axis on the left and Yaw-axis on the right.

This paper presents voice manipulation methods, 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.

2. Voice Manipulation Models

2·1 Problem Statement

Head motions are considered most relevant to the changes in voice. We divide the head motions into two types: the pitch-axis rotation and the yaw-axis rotation shown as Fig. 1. The pitch-axis rotation is vertical whereas yaw-axis rotation is horizontal. We make two assumptions. First, 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 the source filter model [6]. Second, the yaw rotation determines the azimuth direction of

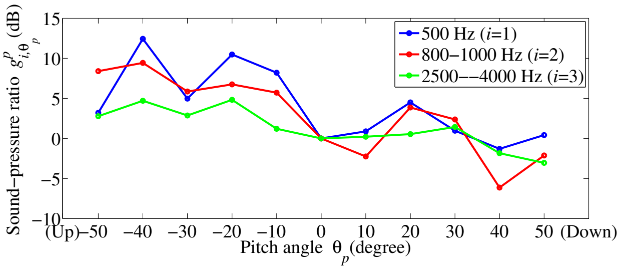

Fig.2: θ_p–gain model for three bands

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

2·2 Pitch Rotation Control

The filter of the vocal tract derived from vertical head motions is denoted by $H(\omega, \theta_p)$. We build a model of $H(\omega, \theta_p)$ by inspecting a human voice for various angles θ_p. Speech signals of a male subject, one of the authors, were recorded with a close-talking microphone in an anechoic chamber. A 10-second-long sweep-tone, with the fundemental frequency ranging $261 - 523$ (Hz), vocalization of vowel /a/ was recorded with the subject's head moving 10° at a time from 50° downward to 50° upward. The subject was instructed to vocalize at the same loudness to emphasize changes in the spectral envelope without changes in the power. The recorded voice signal was then analyzed with one-third octave bands, and sound levels for each band compared to the respective levels at 0° were calculated.

We choose three frequency bands (500, $800 - 1000$, and $2500 - 4000$ (Hz)) to manipulate the voice quality because these bands have formants of most vowels and presents remarkable change in the sound level for varying θ_p. Fig. 2 shows the ratios of sound-pressure level g^{p}_{i,θ_p} in dB to 0° voice for each band and θ_p. The upper suffix p indicates a gain for pitch angle, and i represents the band index. Negative pitch angles indicate facing upward whereas positive ones indicate facing downward.

It is confirmed that the three bands have dominant sound pressure by observing vocalizations of another female subject. We also confirmed that the gain for each band declined similarly to Fig. 2 when the subject faced downward.

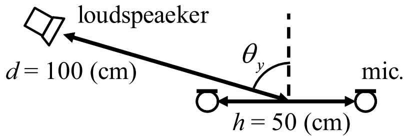

Fig.3: Setup for a left-right balance measurement

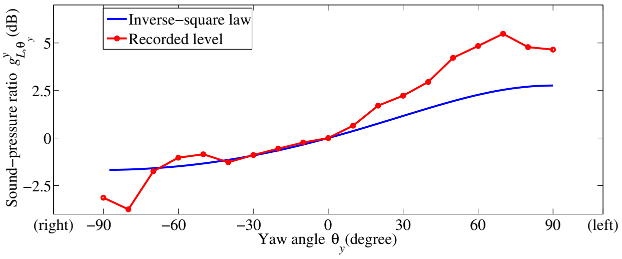

Fig.4: Empirical and theoretical θ_y–gain model for the left channel

2·3 Azimuth Plane Control

We use a pair of loudspeakers and modify the left-right channel balance to present the direction on the azimuth plane. The left-right balance is obtained by measuring impulse responses in an anechoic chamber as shown in Fig. 3. 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. Fig. 4 shows the recorded and simulated sound balance. The simulation is based on the inverse-square law. The y axis represents the sound-pressure ratio compared to 0° 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.

3. Algorithm for Voice Manipulation

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.

3·1 Pitch-axis Modification

In general, the filter $H(\omega, \theta_p)$ is time-variant since the robot may talk while it is moving its head. The spectral envelope modulation is processed in time domain allowing for varying θ_p. First, the voice signal is decomposed into θ_p-dependent and θ_p-independent components. Then, θ_p-dependent components are amplified at each time in accordance with the spectral envelope model. Finally, all components are integrated into a spectral-modulated speech signal.

Decomposition The original signal $x(t)$ is decomposed into $x_i(t)$ and $x_{NULL}(t)$, where $i = 1, 2, 3$ using FIR filters $h_i(t)$ and $h_{NULL}(t)$ with the length 41. The θ_p-dependent components $x_i(t)$ are calculated as

$$x_i(t) = (x * h_i)(t), \tag{1}$$

where $*$ means convolution and h_i are bandpass filters whose center frequencies are determined in Section 2·2. These $x_i(t)$ are amplified in accordance with θ_p. The θ_p-invariant component $x_{NULL}(t)$ is obtained as

$$x_{NULL}(t) = (x * h_{NULL})(t). \tag{2}$$

The waveform $x_{NULL}(t)$ is residual where three signals $x_i(t)$ are taken away from the original $x(t)$. This component is invariable for any θ_p. Therefore, the gain of frequency response for h_{NULL} is zero at 500, 800 – 1000, 2500 – 4000 Hz and one at the other one-third octave band center frequencies.

Amplification The gain for each sample $g_i^p(t)$ is obtained by interpolating the gains shown in Fig. 2 every 10 degrees using $\theta_p(t)$ as shown in Eq. (3). The gain is interpolated with respect to dB.

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

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

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

where g_{i,θ_m}^p 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, $g_i^p(t) = (g_{i,30°}^p + g_{i,40°}^p)/2$. 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 $10^{\frac{g_i^p(t)}{10}}$.

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

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

3·2 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 Fig. 4.

The gain for each channel $g_j^y(t;\theta_y(t))$ $(j = L, R)$ is obtained by interpolating similarly to Eq. (3) as in equations (8) and (9).

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

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

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

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

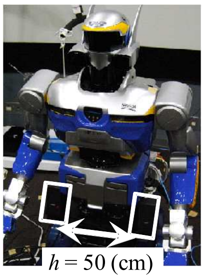

Fig.5: HRP-2 with loudspeakers placed in white boxes

Fig.6: HRP-2 on the left and Robovie-R2 on the right

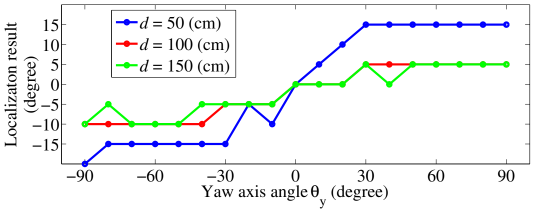

Fig.7: Sound localization result on the azimuth plane

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

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

4. Experimental Evaluation

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.

4·1 Experimental Setup

HRP-2 had a pair of stereo loudspeakers located at its waist as shown in Fig. 5. The space between the speakers was 50 (cm). HRP-2 and Robovie-R2 stood face-to-face with a distance d in a room with moderate reverberation, $RT_{20} = 150$ (ms). Robovie-R2 had eight microphones around its head for sound localization. The experiments were carried out with $d = 50, 100, 150$ (cm) which respectively correspond to intimate, personal, and social distances according to the Proxemics [7]. Speech signals were generated by a speech synthesizer developed by Fujitsu Ltd. The sentence used for this experiment is an excerpt from phonetically balanced sentences in Japanese.

4·2 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 [8]. With this algorithm, Robovie-R2 is able to detect the sound source direction from Robovie's view with its spatial resolution of 5°. Fig. 7 shows the

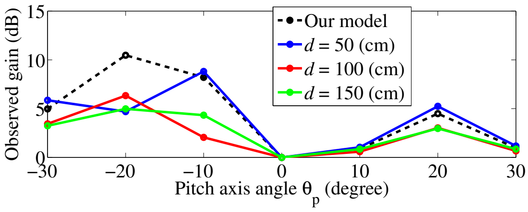

Fig.8: Observed power ratio in 500 Hz band

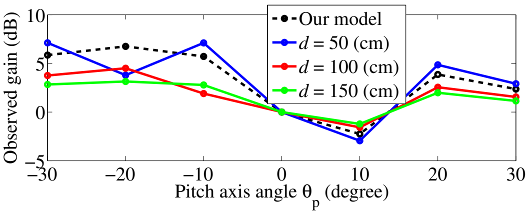

Fig.9: Observed power ratio in 800–1000 Hz band

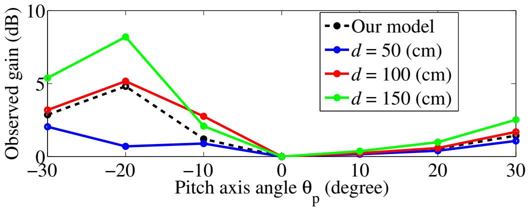

Fig.10: Observed power ratio in 2500–4000 Hz band

results for three distances d. Negative localization angle means left of Robovie's view, meaning HRP-2 was facing rightward. When two robots stands far away, the localization angle diminishes because two loudspeakers were placed only 50 (cm) from each other.

According to Fig. 4, $g^y_{L,50^\circ} - g^y_{R,50^\circ} \approx 6$ (dB) is necessary to let Robovie 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).

4·3 Pitch-axis Directionality

Speech signals from HRP-2 were recorded with the front microphone attached to Robovie. Recorded signals were put through the three band-pass filters in Eq. (1). Ratios of sound-pressure level derived from band-passed signals to those from horizontal vocalization were computed. The pitch angle θ_p ranged from -30° to 30°, which was the range of motion for HRP-2. The results are shown in Figures 8 to 10 for 500, 800–1000, and 2500–4000 Hz bands, respectively.

Observed sound-pressure level ratios conform to our model when Robovie was 50 (cm) away from HRP-2, except for $\theta_p = -20^\circ$. This was caused by a normalization in the amplitude of waveform that was intended to avoid the clipping. Our model has the largest gain for all bands when the head moves upward by 20° and that is prone to cause over-amplification.

Figures 8 and 9 indicate that the effect of pitch-axis movement declined as Robovie draws apart from HRP-2. This was because the power of the reverberation in the room became relatively strong compared to direct speech signal arrival.

By contrast, Fig. 10 shows more power lies in 2500–4000 Hz band as Robovie moved away from HRP-2. This was because the white noise became dominant in that frequency band where the power of a speech signal from a distance was relatively low.

5. Conclusion

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 rotations, pitch and yaw, affect the voice quality independently. The left-right sound-pressure balance for the azimuth direction control is modelled 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. 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 the directionality declines as the distance between two robots becomes larger (150 cm).

Our main future work is top-down modelling of the relationship between vocal tract and physical movements or between vocal band, the source in the source-filter model, and postures. As far as the authors know, psychoacoustic observation about motion-speech relationship, i.e. voice-awareness, has not been investigated. Applicable approaches may be using X-ray imaging or magnetic resonance imaging to visualize vocal organs while a subject is speaking.

References

[1] S. Fujie *et al.* Multi-modal integration for personalized conversation: Towards a humanoid in daily life. In *8th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2008)*, pp. 617–622, Dec. 2008.

[2] R. Dillmann *et al.* Armar II - a learning and cooperative multimodal humanoid robot system. *International Journal of Humanoid Robotics*, Vol. 1, No. 1, pp. 143–155, 2004.

[3] Z. Inanoglu *et al.* Intonation modelling and adaptation for emotional prosody generation. In *Affective Computing and Intelligent Interaction*, pp. 286–293, 2005.

[4] D. Erickson. Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology*, Vol. 26, No. 4, pp. 317–325, 2005.

[5] T. Tasaki *et al.* Distance based dynamic interaction of humanoid robot with multiple people. *Innovations in Applied Artificial Intelligence, LNAI*, Vol. 3533, pp. 111–120, 2005.

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

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

[8] K. Nakadai *et al.* An open source software system for robot audition hark and its evaluation. In *Humanoids 2008*, pp. 561–566, Dec. 2008.