Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Spatial acoustic cues for the auditory perception of speaker's facing direction

Hiroaki Kato, Hironori Takemoto, Ryouichi Nishimura and Parham Mokhtari

National Institute of Information and Communications Technology (NICT)
ATR Bldg., 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

PACS: 43.66.Lj, 43.66.Pn, 43.60.Dh

ABSTRACT

In pursuit of an ultimately realistic human-to-human telecommunication technology, the ability to auditorily perceive the facing direction of a human speaker was explored. Listeners' performance was assessed in an anechoic chamber. A male speaker sat on a pivot chair and spoke a short sentence while facing a direction that was randomly chosen from eight azimuthal angles or three elevation angles. Twelve blindfolded listeners heard the spoken sentence at a distance of either 1.2 or 2.4 m from the speaker and were asked to indicate the speaker's facing direction. In separate sessions, the speaker continuously changed facing angles while speaking and the listeners indicated the perceived direction of horizontal rotation (clockwise or counter-clockwise) or vertical rotation (upward or downward). The overall results showed that the listeners' average response errors were 23.5 degrees for azimuth and 12.9 degrees for elevation. These values were comparable to or better than those obtained in previous studies using a loudspeaker. The average correct-response rates for rotation direction (either horizontal or vertical) were equal to or more than 80%. To identify acoustic cues that have caused the listeners' accurate performance, the acoustic transfer characteristics from the speaker's mouth to the listener's ears were measured by the cross-spectral method. Finer transfer functions were further obtained in a couple of conditions of particular interest by numerical computer simulation using the finite difference time-domain method. The results suggested that major cues included but were not limited to the overall level and spectral tilt for the front-back or up-down judgment, and the interaural level difference for the left-right judgment.

INTRODUCTION

Telecommunications over state-of-the-art audio-visual equipment potentially requires information on where the speaker is facing along with the sound track, especially when using large-scale three-dimensional pictures, due to the apparent closeness between the speaker and listener. Acoustic information at the listener's ears indeed changes with the speaker's facing direction (even when the speaker's position is unchanged) and may seem unnatural if it is incongruent with the visual information. Towards a better understanding of both perceptual and acoustical effects of such information, this paper shows the results of an empirical study designed to measure listeners' ability to identify the facing direction of a human speaker and to estimate the acoustic cues they use.

Different facing directions of a speaker cause different ear input signals for a listener because human vocalization has a nonuniform spatial radiation pattern [1, 2, 3, 4]. A pioneering study in 1939 by Dunn and Farnsworth [1] measured the radiation pattern or directivity of one male speaker on the right hemispheric surfaces (7 radii ranging from 0 to 1 m) at 45-deg angular intervals, and found a strong frequency dependence. About forty years later, a more densely-spaced and precise measurement was done by Moreno and Pfretzschner [2] in both horizontal and median planes. They compared 10 different speakers' directivity patterns and reported a good quantitative agreement among them in terms of the frequency dependence. Chu and Warnock [4] recently published a more comprehensive study that provided 40 speakers' (20 males and 20 females) directivity data on the left hemispheric surface at 15-deg (horizontal) or 20-deg (vertical) intervals for each 1/3-octave frequency band ranging from 160 to 8000 Hz. Their results were generally in good agreement with those of the past reports [1, 2], i.e., within one standard deviation in most directions and frequency bands. Although Studebaker [3] used only sparsely spaced (at 45° intervals) measurement points in the horizontal plane, he targeted "artificial" speakers as well, and concluded that a loudspeaker might exhibit a very different directivity pattern from those of human speakers depending on its dimensions, with a manikin having a moderately different one.

The directivity information of human speakers has been used in the fields of audio-related industry and engineering. In the early stages, directivity patterns were used as reference data by broadcasting or recording engineers in choosing an optimum microphone position, and then to correct the frequency characteristics of recorded speech; the latter usage included preprocessing for automatic speech recognition. More recently, they started to garner intensive attention from applications of automatic talker-orientation extraction in, for example, a video conference situation and a so-called "smart room" environment [5, 6, 7]. These studies demonstrated that their systems successfully exploited human speaker's directivity (either only acoustic information or in combination with video information) to estimate the orientation of the speaker. However, the primary purpose of these systems was not the modeling of human perception, and they were commonly equipped with a microphone array(s) having a fairly larger diameter than a human head, or spreading over an entire room. Therefore, the issue regarding human perception of speaker's facing directions in relation to the usability of directivity information is still an open question.

As already mentioned, advanced audio-visual equipment might be required to record and reproduce speaker's directivity information. However, we cannot determine the necessary specifi-

cations of such information (kinds of acoustic properties and their degrees of accuracy) without knowing the human perceptual characteristics. A couple of studies shared awareness of this problem [8, 9, 10]. Neuhoff *et al.*'s study [8] submitted the notion of the "minimal audible facing angle" or MAFA, which denoted the listeners' ability to perceive a small difference in the facing orientation of a directional sound source. They defined the MAFA as the angle at which listeners achieved a 75% correct response rate in discriminating two different facing directions, and reported the measured MAFAs being 9 and 12 degrees for different distances between the source and the listener. They used a broadband noise source radiated from a small square-shaped loudspeaker placed in a semi-anechoic room. Neuhoff [9] further investigated the listeners' accuracy in identifying the facing direction of a sound source using a small square-shaped loudspeaker placed in a moderately reverberant room, and reported that the average identification errors ranged from 47.0° to 52.5°. Takano *et al.* [10], on the other hand, reported smaller identification errors ranging from 21.3° to 27.9°. They used a loudspeaker embedded in a rigid spherical enclosure whose dimensions resembled a human head (0.17-m diameter) placed in an anechoic room. These three studies commonly suggested the interaural level difference of the listener as a possible cue for the perception of facing directions.

Another commonality among these previous studies was the fact that they used a loudspeaker as the sound source. However, an assessment using a loudspeaker and that using a human speaker would differ in many aspects. As Studebaker pointed out [3], a human speaker's directivity pattern may largely differ from those of the loudspeakers used in these studies. Moreover, the stability or reproducibility of a sound radiated by a loudspeaker is absolutely better than that of human speaker's vocalization. Therefore, one may not directly apply the listeners' performance observed in the loudspeaker cases into human speaker cases, while the estimation of listeners' performance for a living human speaker would be a necessary step towards the evolution of human-to-human telecommunications.

Based on such a background, the present study was designed to first assess how accurately naïve human listeners were able to identify the horizontal or vertical facing direction of a "human speaker" solely by auditory information (speaker's voice), and second to estimate which acoustic cues the listeners exploited.

LISTENING PERFORMANCE

Methods

All procedures strictly complied with the ethical guidelines of NICT.

Participants

Twelve paid adults (7 females and 5 males, ranging in age from 21 to 42) participated in the experiment as listeners. None of them had a history of hearing impairment and all had an air-conduction threshold of 25 dB HL (ISO) or better between 250 Hz and 8,000 Hz in one-octave steps at the time of participation. One male adult (45-year-old) whose native language was Japanese and who had no history of speech impediment participated as a speaker.

Equipment and experimental setup

All experiments were carried out in an anechoic chamber [11], of which the inside dimensions between the tips of absorption wedges were 5.4 m (width) × 4.8 m (depth) × 4.0 m (height). Average background noise level during the experiments was 26 dB SPL (A-weighted), which included noises from a notebook PC and ventilation, and was measured at the position of

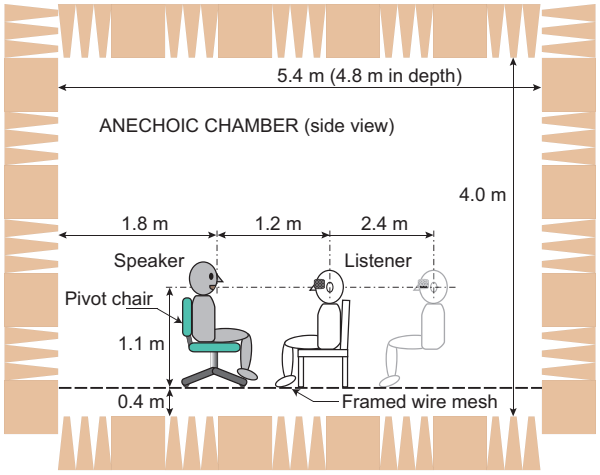


Figure 1: Schematic view of the experimental setup (overview).

the listener's ear by a sound level meter (Type 2231, Brüel & Kjær).

The speaker sat on a pivot chair placed in the chamber. A listener sat directly facing the speaker on an un-rotatable chair placed so that the distance between the speaker's mouth and the listener's ears was either 1.2 or 2.4 m as shown in Fig. 1. The closer distance was chosen so that a subtle airflow caused by the speaker's movement would become well under the listener's detectable level. The listener wore a consumer use but carefully selected blindfold (KC-0746, Kai Corp.). All listeners reported that no motion of the speaker nor pivot chair was visually detectable when they wore the blindfold.

The origin of the speaker's facing azimuthal angle (0°) was set at the direction of the listener, with counterclockwise or clockwise directions being positive and negative angles as viewed from above (Fig. 2, upper panel). Similarly, the origin of facing elevation angle (0°) was set at the horizontal direction, with upper and lower directions being positive and negative angles (Fig. 2, lower panel).

The speaker was able to horizontally rotate by turning the pivot chair using his own feet. The speaker's sitting position was carefully adjusted so that the rotation axis coincided with the tip of his mouth (Fig. 2, upper panel). This was to avoid providing any spatial cue other than the facing direction. If, for example, the speaker rotated around the center of his head, the maximum translational movement of his mouth would be 0.14 m–0.16 m on average [12]. This would correspond to an angular shift of 6.7°–7.6° from a 1.2-m distance, which would probably well exceed the minimal audible angle in the horizontal plane previously obtained at listener's frontal direction (around 1°–4° depending on the source frequency [13, 14, 15]).

To generate a change in vertical facing direction, the speaker simply tilted his head either upward or downward (Fig. 2, lower panel). During the vertical rotation, the maximum translational movement was approximately 0.08 m, which corresponded to an angular shift of 3.8° (1.2-m distance) and was comparable with or smaller than the minimal audible angle in the median plane (on the order of 4° or more depending on different studies [14, 15, 16]).

The speaker was monitored by three 1/4-inch video cameras (QN42H, Elmo) throughout the experimental sessions to check the position of his mouth and the facing direction. The heads of

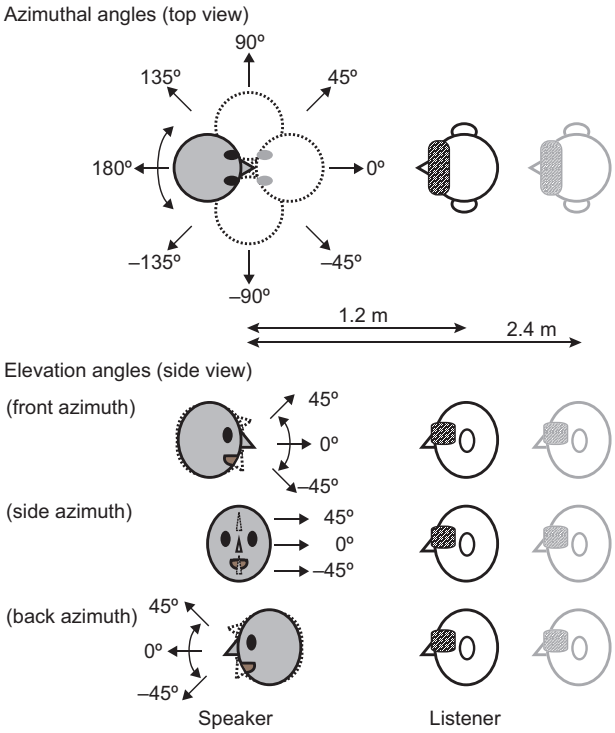


Figure 2: Schematic view of the experimental setup (coordinate system and speaker's movement).

the three cameras were orthogonally coordinated with respect to the speaker's head. Two of them were placed on the walls in front and to the left of the speaker, and the third was attached to the ceiling. The diameter of the camera head was 7 mm and the maximum dimension of the camera holder was 32 mm by 16.5 mm.

Task

Five tasks were undertaken as illustrated in Fig. 3.

A. Azimuthal angle The speaker first faced one of eight azimuthal angles (0°, 45°, 90°, 135°, 180°, –135°, –90°, or –45°) and then spoke a short sentence while holding his face in the initial direction. The listener reported the perceived facing direction of the speaker using one of eight verbal directions, namely: front, right-front, right, right-back, back, left-back, left, and left-front.

B. Horizontal rotation (during speech) A previous study using a loudspeaker suggested that a dynamic change in the facing angle possibly facilitated listener's performance [9]. The present study, therefore, set conditions under which a speaker changed facing directions during speech. The speaker first faced one of the same eight azimuthal angles as the previous task (A), started speaking a sentence, immediately after that rotated the chair (with face and body) either clockwise or counterclockwise by 45 deg (during speech), and then ended speaking the sentence. The listener reported the perceived direction of rotation.

C. Horizontal rotation (during silent interval) To perceive the speaker's rotation direction in the previous task (B), vocalization might not have been necessary 'during' the rotation. That is, a listener might be able to respond correctly solely based on the comparison between the starting and ending directions. The third task was employed to test this possibility. The speaker first faced one of the same eight azimuthal angles as the previous two tasks (A and B), spoke the first half of a sentence, after that rotated the chair (with face and body) either clockwise or counterclockwise by 45 deg (during silence), and then spoke the second half of the sentence. The listener reported the perceived direction of rotation.

D. Elevation angle The speaker first faced one of three elevation angles (0°, 45°, or –45°) and then spoke a short sentence while holding his face in the initial direction. The listener reported the perceived facing direction of the speaker using one of three verbal directions, namely: level, up, and down. The azimuthal facing direction of the speaker was either 0°, –90°, or 180°.

E. Vertical rotation (during speech) The speaker first faced one of the same three azimuthal angles as the previous task (D), started speaking a sentence, immediately after that vertically rotated the head either up or down by 45 deg (during speech), and then ended speaking the sentence. The listener reported the perceived direction of rotation.

Although listeners' heads were not fixed in all the tasks, they were instructed not to move their heads as much as possible. Listeners responded orally in all the tasks. This was because other ways hardly allowed blindfolded listeners to select a response from eight or three alternatives and reliably report it to the experimenter without taking any special training.

Stimulus

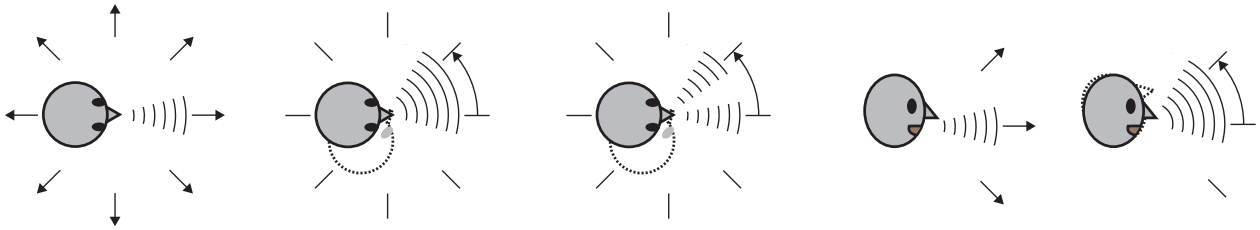
In each trial, the speaker randomly spoke one of the following four sentences that were chosen from a phonetically balanced Japanese sentence set [17]. (English translation follows each alphabetically transcribed Japanese sentence.)

1. *Arayuru genzitu-o subete zibun-no hou-e nezimagetanoda.* (He/she has distorted all (inconvenient) facts towards his/her side.)
2. *Issyukan-bakari nyuyoku-o syuzaisita.* (I/we gathered information about New York city for a week or so.)
3. *Terebigemu-ya pasokon-de gemu-o site asobu.* (He/she plays video games on the TV, PC, etc.)
4. *Uresii hazu-ga yukkuri nete-mo irarenai.* (I first thought it was great (to get a bonus day-off), but I couldn't even sleep in.)

Prior to the experiment, the speaker extensively practiced speaking these sentences and changing his facing directions for three days. The purpose of the practice was to keep the speed and level of all utterances constant as much as possible throughout experimental sessions, as well as to accurately change facing directions. Each training day lasted for three hours, which roughly corresponded to the maximum duration of actual experimental sessions per listener, while being continuously monitored by a stopwatch, a sound level meter, and video cameras.

In the experiment, the speaker explicitly roved the level of speech stimuli; he spoke either louder or softer by approximately 4 dB in randomly selected trials, which corresponded to 25% of all trials. This was to let listeners know that the speaker's voice might fluctuate largely, and that the fluctuation in loudness caused by the speaker's own factors might not provide an effective cue for the correct response.

The durations of the speech stimuli ranged from 3.9 s to 6.0 s. The average of the equivalent sound level of the stimuli was 62.5 dB SPL (A-weighted), which was measured at the center of the (absent) listener's head, a distance 1.2 m from the speaker's mouth.



A. Azimuthal angle

Speaker: Faced one of the eight azimuthal angles and spoke a short sentence.
Listener: Reported the perceived angle.

B. Horizonral rotation (during speech)

S: Horizontally rotated the chair (with face and body) in either direction during speech.
L: Reported the perceived direction of rotation.

C. Horizonral rotation (during silent interval)

S: Spoke the first half of the sentence at one azimuthal angle, rotated silently, and spoke the second half at another angle.
L: Reported the perceived direction of rotation.

D. Elevation angle

S: Faced either of the three elevation angles during speech
L: Reported the perceived angle.

E. Vertical rotation (during speech)

S: Vertically rotated the face either upward or downward during speech.
L: Reported the perceived direction of rotation.

Figure 3: Five different experimental tasks.

Pre-experimental check up

Our low-tech speaker rotation apparatus operated fairly quietly. However, we took several countermeasures against audible noises that may have potentially provided a hint about the speaker's facing direction, just to be safe. Noises accompanied by the rotation of the pivot chair were in most cases below the background noise level at the position of the listener's ear and not audible. Even when they were audible, on rare occasions, possible source positions were limited to the center shaft of the chair, and, therefore, provided no effective information about the facing azimuth. However, the duration of a noise might allow a listener to estimate the amount of rotated angle. To prevent this, a masker noise with a sufficient level and duration was played in inter-trial intervals, during which the chair was rotated towards the starting azimuthal angle for the subsequent trial.

The masker noise was a monaural steady-state pink noise playing through two small loudspeakers (MM-1, Bose) placed on both sides of the speaker's chair facing the listener's direction. The relative level between the two loudspeakers (the balance) was carefully adjusted to the individual listeners so that they localized the fused sound image at the center axis of the pivot chair. The duration of the masker noise was fixed at 5 s, which was chosen as a sufficient length to rotate the chair between two extreme angles. Its average sound level measured at the listener's ear was 58.8 dB SPL and 56.8 dB SPL (A-weighted) under the 1.2-m and 2.4-m distance conditions, respectively. No masker noise was played during trials, in order to maintain a sufficient signal to noise ratio. Although a listener might hear a rotating noise in trials involving the chair rotation (tasks B and C), this should not be regarded as an effective cue because the amount of rotating angle during a trial was fixed at 45 degrees.

In case of incidentally occurring audible noises other than that from the chair's shaft, the experimenter simply excluded the listener's response for that particular trial and appended an identical trial at the end of the same session without notifying the listener.

A preliminary experiment was conducted to test whether or not any effective cue remained other than the speaker's voice after taking these countermeasures. The preliminary experiment set control conditions, in which no speech was presented, in addition to regular with-speech conditions. Results from two listeners showed that correct response rates in the control conditions did not exceed their corresponding chance performances, suggesting that no audible (or inaudible) cue remained in the control without-speech conditions.

Experimental procedures

All listeners took all of the five experimental tasks (A to E). The number of trials in each task was as follows, making a total of 312 trials per listener.

- **A.** 8 azimuthal angles×2 distances×4 sentences = 64 trials
- **B.** 8 azimuthal intervals×2 distances×4 sentences = 64 trials
- **C.** 8 azimuthal intervals×2 distances×4 sentences = 64 trials
- **D.** 3 elevation angles×3 azimuthal angles×2 distances×4 sentences = 72 trials
- **E.** 2 elevation intervals×3 azimuthal angles×2 distances×4 sentences = 48 trials

Literature in spatial hearing has reported that an intensive preliminary practice would considerably affect, in general improve, listener's performance (e.g., [13, 18]). The present study provided listeners no explicit practice to avoid changing their performance from that in their daily lives as much as possible. To assure stability in the listeners' responses instead, randomly selected 4 to 8 dummy trials preceded each session; they were not counted in the final results.

Prior to each trial, the masker noise was played without exception. This noise functioned as an auditory "fixation point," which indicated the precise direction of the speaker since the noise (actually the fused image) had been aligned with the center axis of the speaker. We also expected that the noise would prevent blindfolded listeners from losing their own bearings.

Each listener participated in ten sessions, corresponding to ten conditions resulting from the combination of five tasks and two distances. All listeners took the tasks involving azimuthal angles (A, B, and C) first, and took those involving vertical angles (D and E) second. The other temporal orders between conditions were counterbalanced among listeners. Listeners were allowed to take a break as needed in addition to obligatory inter-session breaks. The whole experimental run including breaks lasted for approximately 135–170 minutes per listener. Before running the second listener of a day, the speaker took a rest for at least three hours.

Results

Overall responses for five tasks

Table 1 summarizes the experimental results for each of the five tasks: the total number of all listeners' responses, the aver-

Table 1: Total number of responses, ratio of correct responses, and average response errors in degree for each task pooled over all listeners with chance performances in the parentheses.

Listener's task		Number of total responses	Percent correct	Average error
A.	Azimuthal angle	768	57% (12.5%)	23.5° (90.0°)
B.	Horizontal rotation (during speech)	768	84% (50%)	—
C.	Horizontal rotation (during silence)	768	86% (50%)	—
D.	Elevation angle	864	74% (33%)	12.9° (40.0°)
E.	Vertical rotation	576	80% (50%)	—

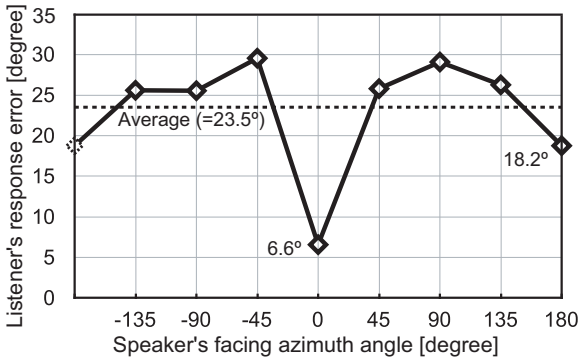


Figure 4: Listeners' accuracy as a function of speaker's facing azimuth.

age ratio of correct responses, and the average response errors in degrees (with each chance performance in parentheses). Differences in the distance between the speaker and listener (1.2 m or 2.4 m) and in the spoken sentence (four sentences) were pooled since they exhibited no critical effects on the overall performances.

An ANOVA was carried out on the correct response ratio (applying an arcsine transformation) of the tasks A, B, and C with the temporal order of tasks within a listener (first or second half), the speech level (normal, louder, or softer), and the initial facing angle of the speaker as main factors (listeners as repetition in the last two factors). The main effect of the speaker's initial facing angle was statistically significant in all the three tasks. Among them, the effect observed in the task A 'azimuthal angle' was remarkable; the next subsection details it. In the tasks B and C, the listeners were more accurate when the speaker was facing his left hemisphere than facing his right hemisphere. Another ANOVA was carried out on the correct response ratio (applying an arcsine transformation) of the tasks C and D with the previously included three factors and the speaker's facing azimuth (0°, –90°, or 180°) as main factors (listeners as repetition in the last three factors). The main effect of speaker's facing azimuth was statistically significant in these two tasks. The listeners were the most accurate when the speaker was facing 180 deg azimuth, i.e., the back direction.

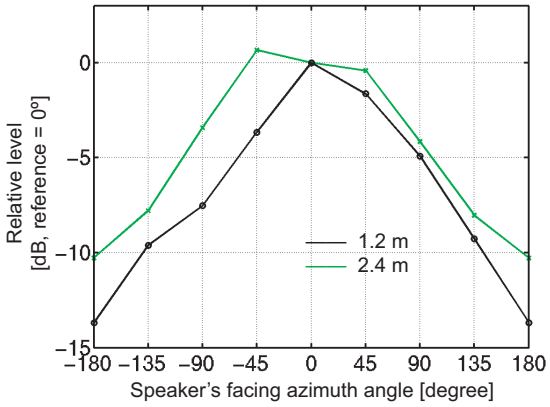


Figure 5: Overall radiation level at the listener's ear as a function of speaker's facing azimuth.

No other main effect was statistically significant.

Azimuthal dependency

As mentioned in the previous subsection, a remarkable effect of the speaker's facing azimuth on the listeners' correct response ratio was observed in the task A. To visualize this effect, the listeners' average response errors were depicted as a function of the speaker's facing azimuth in Fig. 4. Multiple comparisons using the Tukey-Kramer's HSDs (honestly significant differences) indicated that the listeners were the most accurate when the speaker was facing 0 deg azimuth, i.e., the front direction, with the back direction (180 deg azimuth) being the second most accurate.

A similar azimuthal dependency was also observed in a previous study [9] that reported the smallest listeners' errors also when the source azimuth was 0 deg, although the overall averages were different from those of the present study (47.0°–52.5°). A listener's high angular resolution for a speaker facing straight to the listener's direction was also suggested by Neuhoff *et al*'s report [8]. They reported fairly small minimal audible facing angles of 9 and 12 degrees, which were in fact obtained at the frontal source direction. In addition, the average error (23.5°) of the present study was comparable with that reported in another previous study that used a spherical loudspeaker as the source (21.3°–27.9°, [10]). The acoustic causes of the observed azimuthal dependency will be further discussed in a later subsection.

SPATIAL ACOUSTIC CUES

Overall cues for five tasks

To estimate acoustic cues that have caused the listeners' accurate performance, the acoustic transfer characteristics from the speaker's mouth to the listener's ears were measured by the cross-spectral method using the speaker's own voice as the test signal. The same speaker who had participated in the listening experiment wore a 1/4-inch microphone (Type 4951, Brüel & Kjær) held close to the mouth (65-mm right of the center of the mouth and 15-mm forward of the lips). A listener who had participated in the preliminary experiment wore binaural microphones (SP-TFB-2, Sound Professionals) at the entrances of the right and left ear canals. Following Nukina and Kawahara's methodology [19], the speaker spoke sustained five vowels /a/, /i/, /u/, /e/, and /o/ twice. Phonation of each vowel lasted about 6 s and was quasi-sinusoidally frequency-modulated at about 0.5 Hz with the carrier frequency ranging approximately

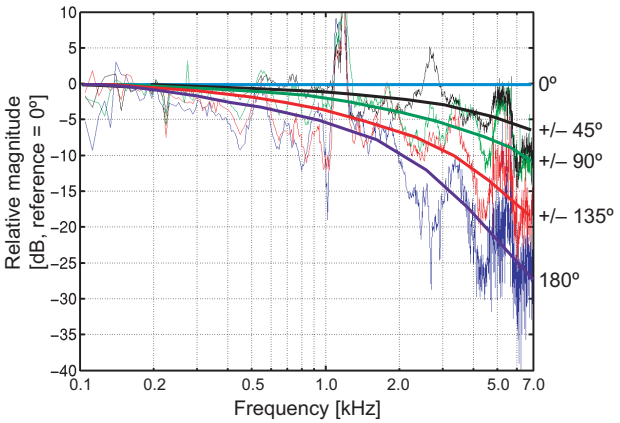


Figure 6: Transfer functions between speaker and listener. Different facing azimuth of the speaker are expressed by different colors.

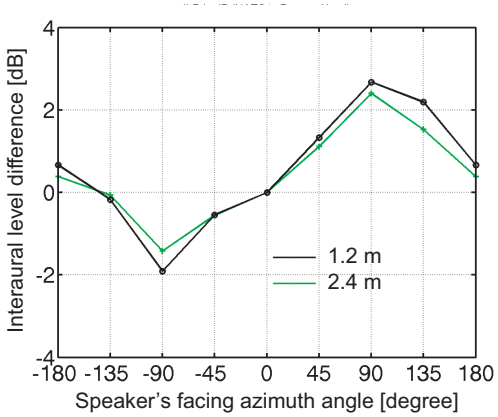


Figure 7: Listeners' interaural level difference as a function of speaker's facing azimuth.

from 110 Hz (A2) to 220 Hz (A3). The speaker also spoke the same four sentences used in the listening experiment three times. Acoustic measurements using these speech signals were repeated for each of all combinations of the speaker's facing directions and listening positions tested in the listening experiment, resulting in an impulse response for each of the experimental conditions.

Consequently, we examined possible acoustic cues derived from the impulse responses obtained above. First, we investigated two potential cues for listeners' judgment along the front-back dimension. Figure 5 shows the relationship between the overall sound level transfered to the listener's ears and the speaker's facing azimuth; relative values to that of 0-deg azimuth (front direction) are depicted. Differences between right and left ears are pooled. As shown in the figure, the maximum change in the overall level reached approximately 13 dB or 10 dB in the 1.2-m or 2.4-m distance conditions, respectively, which largely exceeds the detection level (0.4–1.2 dB in optimum conditions [20]). The overall sound level is known as a primary perceptual cue for the distance of a sound source [21], which shares the same dimension as front-back judgments in the present study. It would be, accordingly, tenable that listeners of this study used a difference in the overall level as a cue for the change in speaker's facing directions along the front-back dimension. However, the overall level cue may not be fully reliable when

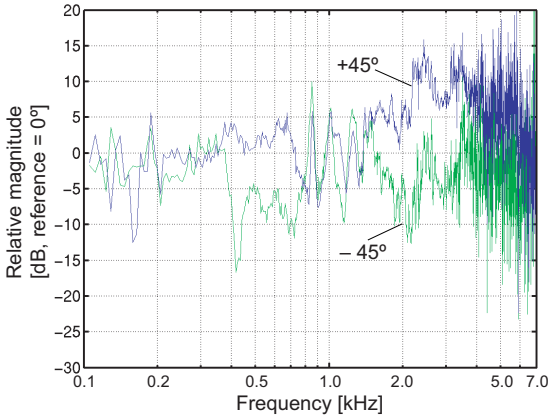


Figure 8: Transfer functions between speaker and listener. Different facing elevations of the speaker are expressed by different colors. The facing azimuth is 180° (=back direction).

the level of the sound source itself is unstable. Since the source level of the present listening experiment was randomly roved by the speaker, the listeners were not able to solely rely on it.

Another potential cue for the front-back judgment was the change in spectral tilts with the speaker's facing azimuth as shown in Fig. 6. Relative values to that of 0-deg azimuth (front direction) are depicted. Presented frequency range is limited from 0.1 to 7.0 kHz, during which we obtained a sufficient gain in the speaker's voice that enabled reliable calculation of a transfer function. Differences between right and left ears are pooled. Differences between right and left hemispheres are also pooled for simplification. A parabolic regression curve is superimposed for each transfer function to indicate general tendency in the spectral tilt. Different azimuths are expressed by different colors. As shown in the figure, the downward spectral tilt from low to high frequency became steeper as the absolute value of the speaker's facing azimuth increased (=rotated from front to back). This correlation is in principle independent of the overall level or perceived loudness of a source. Furthermore, similar spectral changes caused primarily by the occlusion of the source by head and torso had been reported to be effectively used in monaural sound localization [22]. Therefore, the observed change in spectral tilt was likely to be used as a cue for the front-back difference in the speaker's facing direction.

Second, we investigated a potential cue for listeners' judgment along the left-right dimension. Figure 7 shows the relationship between the listener's interaural level difference (ILD) and the speaker's facing azimuth; relative values to that of 0-deg azimuth (front direction) are depicted. A positive value means that the gain in the right ear was larger than that in the left ear. A clear tendency according to the speaker's facing direction was visible, i.e., a positive ILD was observed when the speaker was facing the right hemisphere (in listener's view) and vice versa. This observation suggested that the listeners possibly exploited the ILD as a cue for their judgments along the left-right dimension. However, the amount of the observed ILD was about 2 dB or less, which might not be sufficiently large compared to the detection threshold (0.5–0.8 dB [23, 24]). Therefore, additional acoustic cues should be explored to fully understand the listener's accurate performance along the left-right dimension. A band-specific ILD could be an effective cue, although further study on this is beyond the scope of this paper.

With regard to potential cues for the up-down judgment, we

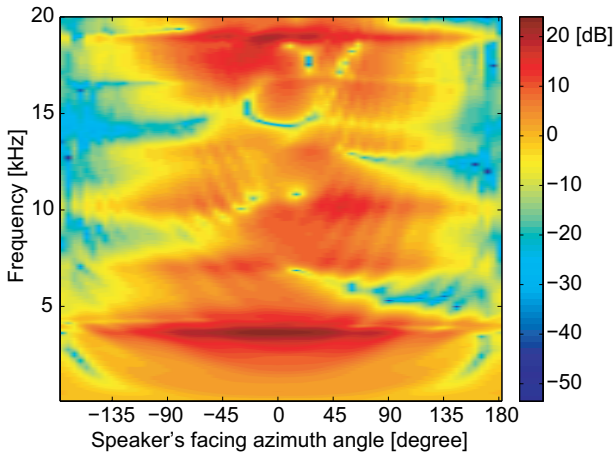


Figure 9: Frequency characteristics of acoustic radiation from speaker's mouth (simulated).

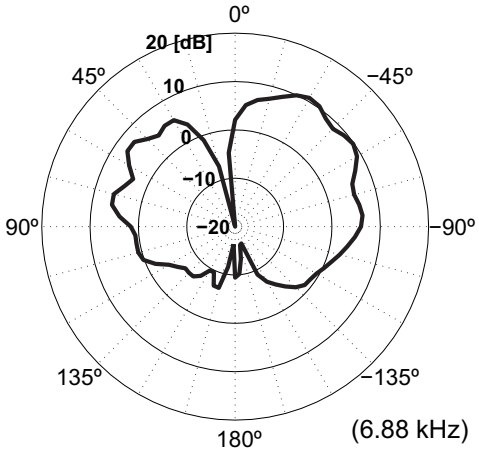


Figure 10: Acoustic directivity pattern of the human speaker at a specific frequency band (centered at 6.88 kHz), numerically calculated at 5-deg intervals at a distance of 1.2 m from the speaker's mouth.

can consider them in a similar way to those for the front-back judgment. The overall level was in general larger at a higher elevation and the spectral shape systematically changed as exemplified in Fig. 8. Presented frequency range is also limited from 0.1 to 7.0 kHz.

Azimuthal dependency

This subsection explored the cause of azimuthal dependency observed in the listeners' performances for the task A 'azimuthal angle.' As shown earlier in Fig. 4, the listeners made particularly small errors (6.6° on average) when the speaker was facing toward the listener's direction (0°). The second smallest errors (18.2° on average) were marked when the speaker was facing toward the back direction (180°). At least two possible causes should be considered. First, the sound produced at 0 deg was acoustically unique, and therefore, listeners were able to easily distinguish it from those produced at other angles. Second, listeners were sensitive to 0-deg facing azimuth. In other words, we are particularly sensitive to whether or not the speaker is turned toward our direction.

The first possibility would be in part supported. The acoustic difference of input signals between listener's left and right ears was probably minimal when the speaker's facing azimuth was 0 deg. In that sense, the sounds at 0 deg had an acoustically unique attribute that might facilitate listener's identification. However, this was also the case with the sounds at 180 deg. This fact does account for the listeners' good performance at 180 deg but does not account for their far better performance at 0 deg.

For a closer test of the first possibility, the acoustic properties measured in the previous subsection were further examined. More specifically, absolute differences between two successive azimuths were calculated and compared to each other. If any difference between the frontal (0°) and adjacent (±45°) azimuths was particularly larger than the others, then it might account for the prominent performance observed at 0 deg. However, this was not found to be the case for the measured data shown in Figs. 5 to 7. A finer comparison in each frequency band would merit consideration because there had been evidence for a listener exploiting a change in a specific frequency band in many aspects of spatial hearing [25, 26, 27]. We therefore performed detailed comparisons frequency by frequency, based on transfer functions numerically obtained by computer simulation. The measured data were not used because their stability across different azimuths might not be valid for such a fine comparison.

The three dimensional anatomical data of the speaker's head was first obtained by volumetric magnetic resonance imaging (ECLIPSE 1.5T, Shimadzu-Marconi, operated at ATR Brain Activity Imaging Center), with a 1.1 mm spatial resolution. The sound propagation characteristics from the speaker's mouth were then calculated using the finite difference time-domain method [28] combined with the near to far field transformation algorithm [29], which dramatically reduced the number of computational steps without deteriorating the accuracy. Figure 9 shows the calculated frequency characteristics at 1.2-m distance excited by a Gaussian pulse placed at the center of the speaker's mouth, as a function of speaker's facing azimuth at 5-deg intervals.

A thorough comparison across the facing azimuths in every 40-Hz frequency bin of the above data revealed a small number of peculiar frequency bands. They exhibited a significant disparity between the frontal (0°) and adjacent (±°45) azimuths in the speaker's directivity patterns. An extreme example is shown in Fig. 10. We, however, have to be prudent because the existence of these frequency bands itself does not necessarily mean the listeners actually used them as acoustic cues.

Since the first possibility was only partly supported as shown above, we should not reject the second one; they were not even mutually exclusive. The notion that a listener is particularly sensitive to whether or not a speaker is turned toward his/her direction would be interesting from psychological and biological points of view. An empirical investigation should address this issue in the future.

CONCLUSIONS

This paper first assessed a listener's ability to identify the facing direction or rotating direction of a human speaker and, second, explored possible acoustic cues that the listener exploited by means of both actual acoustic measurements and numerical computer simulations. Tentative conclusions read as follows:

1. Overall identification accuracy was 23.5 degrees for azimuthal angles and 12.9 degrees for elevation angles.
2. Listeners' accuracy depended on the speaker's facing azimuth; the best performance was achieved when the speaker was facing the listener's direction.

3. Correct identification rate of rotating direction was 84–86% for horizontal rotations and 80% for vertical rotations.
4. Major acoustic cues for the front-back or up-down judgment were suggested as the overall level and spectral tilt.
5. An acoustic cue for the left-right judgment was suggested as the interaural level difference.

These results provided a clue as to the design of a more realistic sound track in multi-modal telecommunications.

ACKNOWLEDGMENTS

The authors thank Tatsuya Kitamura for his support in processing acoustic data and providing the initial impetus for this study, and Hideki Kawahara and Masanori Morise for their instruction on a method for calculating transfer functions using the speaker's own voice.

REFERENCES

[1] Dunn, H. K. and Farnsworth, D. W. (**1939**). "Exploration of pressure field around the human head during speech," J. Acoust. Soc. Am. **10**, 184–199.

[2] Moreno, A. and Pfretzschner, J. (**1978**), "Human head directivity in speech emission: A new approach," Acoust. Letters **1**, 78–84.

[3] Studebaker, G. A. (**1985**). "Directivity of the human vocal source in the horizontal plane," Ear and Hearing **6**, 315–319.

[4] Chu, W. T. and Warnock, A. C. (**2002**). "Detailed directivity of sound fields around human talkers," Tech. Rep. IRC-RR-144, Institute for Research in Construction, National Research Council Canada.

[5] Canton-Ferrer, C., Segura, C., Cacas, J. R., Pardàs, M., and Hernando, J. (**2008**). "Audiovisual head orientation estimation with particle filtering in multisensor scenarios," EURASIP Journal of Advances in Signal Processing, vol. 2008, no. 276846.

[6] Nakano, A. Y., Nakagawa, S., and Yamamoto, K. (**2009**). "Automatic estimation of position and orientation of an acoustic source by a microphone array network," J. Acoust. Soc. Am. **126**, 3084–3094.

[7] Levi, A. and Silverman, H. (**2010**). "A robust method to extract talker azimuth orientation using a large-aperture microphone array," IEEE Trans. Audio, Speech, Lang. Process. **18**, 277–285.

[8] Neuhoff, J. G., Rodstrom, M.-A., and Vaidya, T. (**2001**). "The audible facing angle," Acoust. Res. Letters Online (ARLO) **2**, 109–114.

[9] Neuhoff, J. G. (**2003**). "Twist and shout: Audible facing angles and dynamic rotation," Ecolog. Psychol. **15**, 335–351.

[10] Takano, H., Hokari, H., Shimada, S., and Sugiyama, K. (**2005**). "A study on a perception of the speech-direction," Tech. Rep. Inst. Electron. Inform. Comm. Eng. **105** (348), 37–42. (in Japanese)

[11] Tohkura, Y. (**1989**). "Introduction to the ATR new laboratories in the Kansai Science City," J. Acoust. Soc. Jpn. **45**, 493–494. (in Japanese)

[12] Kouchi, M. (**2005**). "AIST Anthropometric Database 1991–92 Manual," Digital Human Research Center, AIST, http://riodb.ibase.aist.go.jp/dhbodydb/91-92/manual/chapter-6.pdf, p. 30. (in Japanese)

[13] Mills, A. W. (**1958**). "On the minimum audible angle," J. Acoust. Soc. Am. **30**, 237–246.

[14] Wightman, F. L. and Kistler, D. J. (**1989**). "Headphone simulation of free-field listening. II: Psychophysical validation," J. Acoust. Soc. Am. **85**, 866–878.

[15] Makous, J. C. and Middlebrooks, J. C. (**1990**). "Two-dimensional sound localization by human listeners," J. Acoust. Soc. Am. **87**, 2188–2200.

[16] Blauert, J. (**1997**). *Spatial Hearing: The Psychophysics of Human Sound Localization (Rev. ed.)* (MIT, Cambridge, MA), p. 44.

[17] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., and Shikano, K. (**1990**). "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Commun. **9**, 357–363.

[18] Middlebrooks, J. C. and Green, D. M. (**1991**), "Sound Localization by human listeners," Ann. Rev. Psychol. **42**, 135–159.

[19] Nukina, M. and Kawahara, H. (**2003**). "Transfer characteristics of speech sounds around speaker's head," J. Acoust. Soc. Jpn. **59**, 256–260. (in Japanese)

[20] Miller, G. A. (**1947**). "Sensitivity to change in the intensity of white noise and its relation to masking and loudness," J. Acoust. Soc. Am. **19**, 609–619.

[21] von Békésy, G. (**1949**). "The moon illusion and similar auditory phenomena," Am. J. Psychol. **62**, 540–552.

[22] Perrott, D. R. and Elfner, L. F. (**1968**), "Monaural localization," J. Audit. Res. **8**, 185–193.

[23] Rowland, R. C. and Tobias, J. V. (**1967**). "Interaural intensity difference limen," J. Speech Hear. Res. **10**, 745–756.

[24] Yost, W. A. and Dye, R. H. (**1988**). "Discrimination of interaural differences of levels as a function of frequency," J. Acoust. Soc. Am. **83**, 1846–1851.

[25] Hebrank, J. and Wright, D. (**1974**). "Spectral cues used in the localization of sound sources on the median plane," J. Acoust. Soc. Am. **56**, 1829–1834.

[26] Middlebrooks, J. C. (**1992**). "Narrow-band sound localization related to external ear acoustics," J. Acoust. Soc. Am. **92**, 2607–2624.

[27] Shub, D. E., Carr, S. P., Kong, Y., and Colburn, H. S. (**2008**). "Discrimination and identification of azimuth using spectral shape," J. Acoust. Soc. Am. **124**, 3132–3141.

[28] Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (**2007**). "Comparison of simulated and measured HRTFs: FDTD simulation using MRI head data," in *Proc. Audio Eng. Soc. 123rd Conv.* (New York, USA), paper 7240, 12 pp.

[29] Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (**2008**). "Efficient computation of HRTFs at any distance by FDTD simulation with near to far field transformation," in *Proc. Fall Meet. Acoust. Soc. Jpn.* (Fukuoka, Japan), pp. 611–614.