International Conference on Control, Automation and Systems 2007
Oct. 17-20, 2007 in COEX, Seoul, Korea

Sound Direction Estimation using Artificial Ear

Sungmok Hwang[1], Youngjin Park[2] and Younsik Park[3]

[1] Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea
(Tel : +82-42-869-3076; E-mail: tjdahr78@kaist.ac.kr)
[2] Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea
(Tel : +82-42-869-3036; E-mail: yjpark@kaist.ac.kr)
[3] Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea
(Tel : +82-42-869-3020; E-mail: yspark@kaist.ac.kr)

Abstract: We propose a novel design of artificial robot ear and sound direction estimation method using the measured two output signals only. The spectral features in head-related transfer functions and in interaural transfer functions are distinctive across voice frequency band. Thus, these features provide effective sound cues to estimate sound direction using the measured two output signals. Bilateral asymmetry of microphone positions can enhance the estimation performance even in the median plane where interaural differences are vanished. Sound direction is estimated from interaural time difference and correlation between the log magnitudes of interaural transfer functions. The feasibility and the estimation performance of the designed artificial ear and the estimation method are verified in a real environment. In the experiment, we confirm that robots with the proposed artificial ear can find the direction of user from two output signals only with reasonable accuracy.

Keywords: Sound direction estimation; head-related transfer function; interaural transfer function; artificial ear.

1. INTRODUCTION

Sound direction estimation is to estimate the direction of sound source using the measurements of the acoustic signals by a set of microphones. A robot operating in a household environment should detect various sound events and take notice of them to achieve robust recognition and interaction with user. Therefore, sound direction estimation is one of the most critical building blocks in robot technology. In the last few decades, many different algorithms for sound direction estimation have been developed. Most of them depend on inter-channel time difference or inter-channel level difference. However, when inter-channel differences obtained from two output signals are used, only 1-D localization is possible because there are many source points in 3-D space sharing the same inter-channel differences, and this set of points is often called the cone of confusion [1]. That is, inter-channel differences provide the information for the perception of sounds originating from the horizontal plane, which includes both microphones. Thus, many localization methods rely on an array of more than two microphones to estimate azimuth and elevation angles of sound source [2-3]. Several methods based on the integration of vision and audio information have also been developed [4-5]. However, those methods require additional imaging systems and signal processing algorithms.

Human beings, however, can perceive the sound direction and distinguish whether a sound is coming from front or rear, above or below with only two ears. It has been discovered that human auditory system mainly relies on the primary sound cues, which include interaural time difference (ITD), interaural level difference (ILD), and spectral modification due to the complex shape of pinna folds, to perceive the sound direction [1,6-7]. These sound cues are contained in head-related transfer function (HRTF), which is an acoustic transfer function from a sound source to a listener's eardrum [8]. It is known that interaural differences mainly contribute to the lateral perception and spectral modification do to the front-back discrimination and the vertical perception [9-10]. Shaw and Teranishi described that the spectral features like peaks and notches in HRTFs are entailed by the direction dependent acoustic filtering due to the pinna [11-13]. Among the many structures in pinna, the concha is mainly responsible to the vertical perception and the front-back discrimination [10,14]. Hebrank and Wright hypothesized that reflections from the posterior wall of the concha alone may be responsible for the observed notch in the median plane, and they proposed a simple reflection model to derive the spectral notches [6]. Lopez-Poveda and Meddis proposed a diffraction/reflection model of the concha wall to derive the notch frequencies for elevated sources on vertical planes, and they reproduced the spectral notches more accurately [15].

If an artificial robot ear, which can provides proper sound cues as human pinna do, is designed, it is expected that sound direction in 3-D space can be estimated from two output signals only. In our previous study, an artificial ear having a similar shape to the human concha was designed, and HRTFs were measured in an anechoic chamber with respect to the change of source position in 3-D space [16]. Distinctive spectral features varying according to sound direction were shown in the measured HRTFs. The spectral features, however, were distributed at higher frequency band than voice frequency band. The main sound source is human voice signals, thus the spectral features should be distinctive across voice frequency band. Thus, this paper addresses new designs of artificial ear and shows

* Portions of this work were presented at the 16th IEEE International Symposium on Robot and Human Interactive Communication in August 2007.

978-89-950038-6-2-98560/07/$15 ⓒICROS

that spectral features of the new design are distinctive across voice frequency band. Sound direction estimation method is also proposed, and the feasibility and the estimation performance of the design and method are verified from an experiment carried out in a real environment.

2. ARTIFICIAL EAR DESIGN AND ITF MEASUREMENT

2.1 Ear design

An artificial ear for robots mimicking the human concha was designed, and mock-ups of the artificial ear and the robot head were manufactured as depicted in Fig. 1.

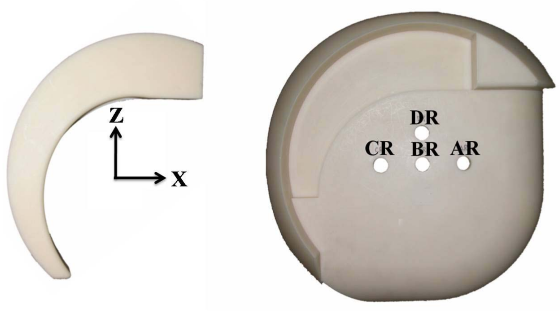

Fig. 1 Designed ear and microphone positions.

The basic assumption guiding our design is that the direction-dependent spectral features are closely related with the distance from a meatus entrance to a posterior wall of concha. Especially, the spectral notch is caused by cancellation between the direct wave reaching the meatus entrance and the reflected wave from the concha posterior wall. This assumption was verified by the previous studies [6,15-16]. Thus, we designed a spiral-shaped artificial ear so as to vary the distance from a microphone to a concha posterior wall according to the sound direction.

2.2 Microphone positions

When two microphones are placed at symmetric positions in the head as like human's left and right ears, interaural differences for the median sources are vanished and spectral modifications of two ear outputs are the same. However, robot should detect the elevation angle of a sound source even in the median plane. The bilateral asymmetry of ear shape and/or microphone position can alleviate this problem. Thus, in this study, we consider four microphone positions at each ear. In Fig. 1, A, B, C, and D indicate the microphone positions, and R (L) indicates the right (left) ear.

2.3 ITF measurement

The input signal is unknown in many practical situations. Therefore, the interaural transfer function (ITF) fusing the two output signal characteristics is a useful and practical cue for sound direction estimation because ITF contains all information about ITD, ILD, and spectral modifications. An H1 or H2 estimator can be used to obtain the ITF from the measured two output signals [17]. The ITF can also be written as the ratio of the HRTFs at the two ears as

$$ITF(\theta,\phi,f)=\frac{H_{ear1}(\theta,\phi,f)}{H_{ear2}(\theta,\phi,f)} \quad (1)$$

where H is the empirical HRTF. The ITF at each frequency (f) varies with the azimuth and elevation angles and includes information regarding interaural differences and spectral modification.

To investigate the distribution of spectral features, interaural transfer functions (ITFs) were measured in an anechoic chamber as depicted in the left figure of Fig. 2. The artificial ear was mounted on the head and microphones (B&K Type 4130) were located in the artificial ear. A white noise with a bandwidth covering the frequency range from 200Hz to 5000 kHz was used as the general input to speaker. This input signal together with the resulting output signal from the microphone were each collected for a sampling duration of 1.5 second at a sampling rate of 16 kHz. HRTFs for 357 source positions in 3-D space were measured in the vertical-polar coordinates as depict in the right figure of Fig. 2. And then, ITFs were obtained by eq. (1).

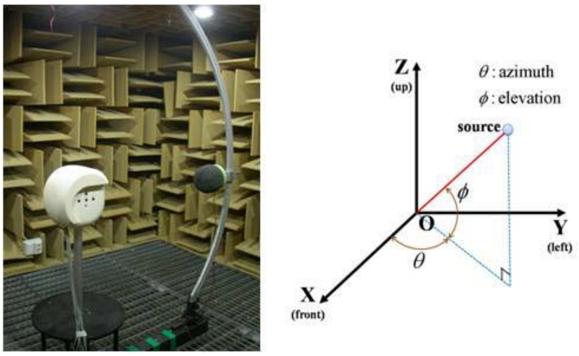

Fig. 2 HRTFs measurement setup (left) and vertical-polar coordinates (right).

The log magnitude of median ITFs in voice frequency band are shown in Fig. 3. The description marked at the top of each panel indicates the combination of ear positions with bilateral asymmetry. For example, the description "BR/CL" indicates that the signals measured at right B and left C positions are considered as output and input for the ITF estimation, respectively. The first panel in Fig. 3 shows the distinctive spectral features whose center frequencies vary according to the change of sound direction in the frontal region where the elevation of source is from -30° to 90°. Especially, the spectral features move monotonically with respect to the varying of the elevation angle of a sound source, and this can be interpreted as the distance from a microphone to a concha posterior wall varies monotonically with change of the elevation angle. The spectral features, however, disappear in the rear region where the elevation is from 90° to 210° because the microphone is hidden by the concha wall and the reflection from the concha wall is not occurred. The results for other combinations, which are shown in the second and third panels, are similar with those in the first panel. These characteristics are almost the same as those in the human HRTFs or ITFs [7]. Therefore, these spectral features provide effective cue for the elevation estimation and the font-back discrimination. Thus, if we build the database containing the spectral features in ITF obtained in an anechoic chamber, then the sound direction can be estimated by comparing the spectral features in the database with those in the ITF obtained from the measured two output signals.

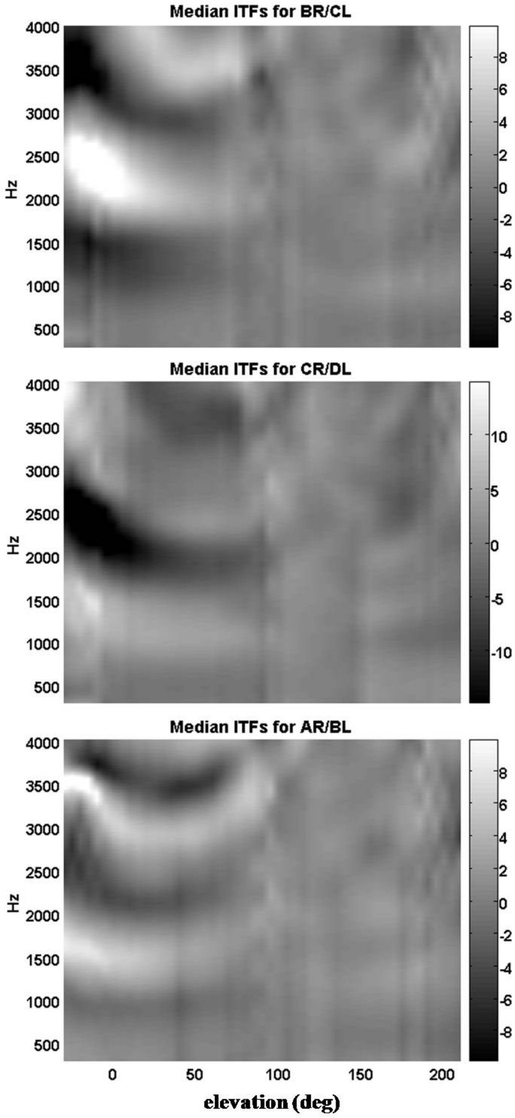

Fig. 3 The log magnitude of median ITFs for different combination of the bilateral asymmetry microphone positions in voice frequency band.

3. ESTIMATION METHOD

In this research, we use the ITD and the log magnitude of ITF to estimate sound direction from the two measure output signals only. ITD is obtained from the Generalized Cross-Correlation method (GCC) [18]. Especially, we use the phase transform (PHAT) which is one of the most successful formulations of GCC. ITF is obtained from H1 estimator. However, the obtained ITF in a real environment has many microscopic fluctuations because the output signals are contaminated by noises such as measurement noise, reflections from furniture and walls. Thus, these microscopic fluctuations should be removed to extract the essential spectral features. We smoothed the ITF by the Gaussian filter as

$$\widetilde{M_e}(k) = \sum_{n=-n_1}^{n_1} M_e(k+n)G(n), \tag{2}$$

where k and n indicate discrete frequency. $M_e(k)$ is the log magnitude of ITF, and $G(n)$ is a Gaussian filter defined as

$$G(n) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-n^2}{2\sigma^2}}. \tag{3}$$

n_1 in eq. (2) and σ in eq. (3) are parameters to determine the filter. An example of the smoothing process is shown in Fig. 4.

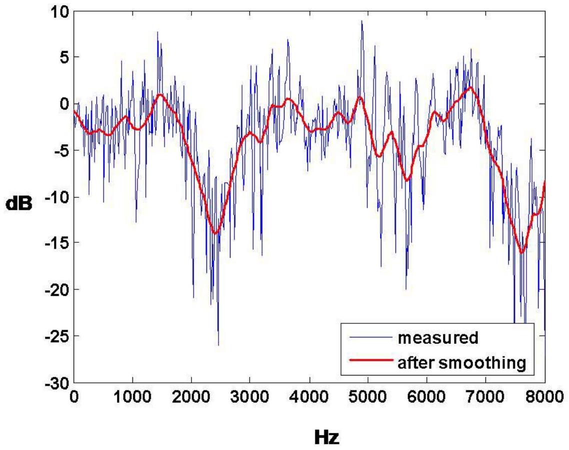

Fig. 4 Before and after smoothing of ITF by Gaussian filter.

Then, the correlations between the smoothed ITF and ones in the database, which were constructed in an anechoic chamber, are computed in the interest frequency band. The sound direction is determined by finding the ITF in database that maximizes the correlation. To reduce the computation load, the only ITF data sharing the same ITD with the measured one by the PHAT are used for computation of correlation. These estimation procedures are summarized in Fig. 5.

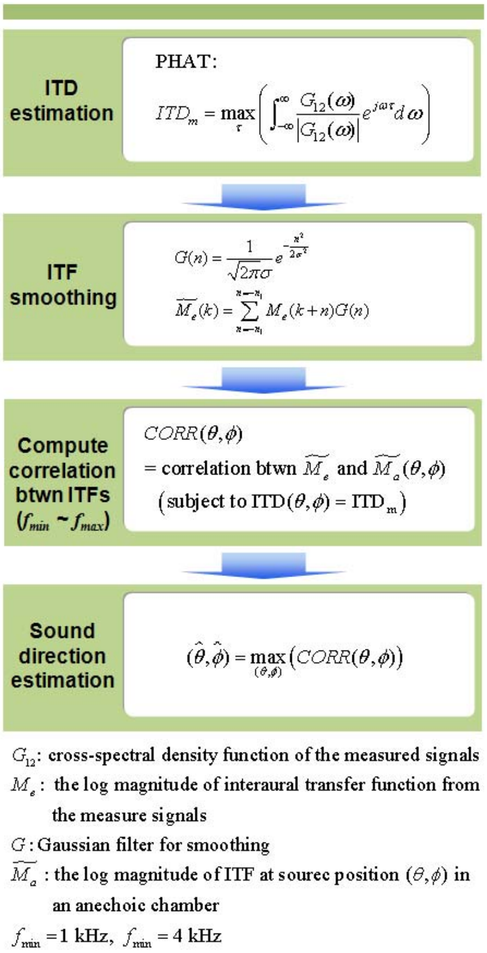
ITD estimation
PHAT:
$ITD_m = \max_{\tau}\left(\int_{-\infty}^{\infty}\frac{G_{12}(\omega)}{|G_{12}(\omega)|}e^{j\omega\tau}d\omega\right)$
ITF smoothing
$G(n) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{n^2}{2\sigma^2}}$
$\widetilde{M_e}(k) = \sum_{n=-n_1}^{n=n_1} M_e(k+n)G(n)$
Compute correlation btwn ITFs (f_{min} ~ f_{max})
$CORR(\theta,\phi)$
= correlation btwn $\widetilde{M_e}$ and $\widetilde{M_a}(\theta,\phi)$
(subject to $\mathrm{ITD}(\theta,\phi) = \mathrm{ITD}_m$)
Sound direction estimation
$(\hat{\theta},\hat{\phi}) = \max_{(\theta,\phi)}\left(CORR(\theta,\phi)\right)$
G_{12}: cross-spectral density function of the measured signals
M_e : the log magnitude of interaural transfer function from the measure signals
G : Gaussian filter for smoothing
$\widetilde{M_a}$: the log magnitude of ITF at sourec position (θ,ϕ) in an anechoic chamber
f_{min} = 1 kHz, f_{min} = 4 kHz

Fig. 5 Flow chart for sound direction estimation.

As shown in Fig. 3, the direction-dependent spectral features are not shown in the rear region. Thus, the proposed ear design has limitation that the elevation of rear source cannot be estimated. However, the front-back discrimination can be achieved because the difference between ITFs of the frontal source and rear source is distinctive. Therefore, in this research, sound direction is estimated for the frontal sources and the front-back discrimination is carried out for the rear sources only.

4. EXPERIMENT

In a real environment, the output signals can be contaminated by noises. Therefore, the measured ITF from two measured output signals can be inaccurate, and the spectral features can be distorted by noises. Thus, to investigate the feasibility and the estimation performance of the proposed ear design and estimation method, an experiment is carried out in a real environment. Two microphones were placed at the left B and the right C positions in Fig. 1. The size of room was 7 m × 7 m × 3 m. Background noise level of the room is approximately 55 dB (A-weighted SPL) and SNR is 20 dB. A male speech signal lasting approximately 0.8 sec was collected in an anechoic chamber, and this was presented by a horn driver speaker in a real room. The distance from the head center to the speaker was set to be 1 m. The azimuth angle of speaker varied from -180° to 180° with 30° intervals and the elevation angle varied from -30° to 30° with 15° intervals. 10 tests were conducted at each source position. To represent the estimation performance, we used two indexes, the rate of front-back discrimination and the estimation error which is defined by the angle (0° ~ 180°) between the unit vectors corresponding to the actual source direction and the estimated source direction. As mentioned in section 3, the front-back discrimination was carried out for the rear sources whereas both the front-back discrimination and the direction estimation were carried out for the frontal sources. On all the tests, the rate of front-back discrimination was 99.6% and 73.7% for the frontal sources and the rear sources, respectively. That is, the frontal sources can be estimated that they is placed in the front region with almost perfect accuracy, however the rear sources can be misestimated as the frontal sources with the rate of 26%. Mean of the estimation errors at each source position in the frontal region is summarized in Table 1. Note that the error increases as the source leans toward one ear (θ=±90°). The reason why the performance degrades for these sources is that the scattering and diffraction effects for a hidden ear due to the head are most dominant. Thus, the power of true output signal is low and SNR is degraded. Although there exist several positions where the error is large, sound direction can be estimated with 15.32° of error on average.

Table 1 Direction estimation errors for the frontal sources.

(degrees)

θ \ ϕ	-30	-15	0	15	30	Ave.
-90	70.15	7.52	10.82	6.91	19.12	**22.90**
-60	0.00	6.91	13.48	9.74	7.68	**7.56**
-30	4.33	17.38	5.00	0.00	9.10	**7.16**
0	22.35	9.66	15.00	10.00	0.00	**11.40**
30	6.68	23.64	17.96	30.42	6.32	**17.00**
60	14.82	6.78	13.61	14.03	35.00	**16.85**
90	36.29	5.00	11.17	25.49	43.77	**24.34**
Ave.	**22.09**	**10.98**	**12.43**	**13.80**	**17.28**	**15.32**

5. SUMMARY AND CONCLUDIONS

We proposed a design of an artificial ear, which mimics the human concha, for robots to estimate the sound direction from measured two output signals only. The estimation method is based on ITD and spectral features in ITF. We used the bilateral asymmetry of the ear positions to enhance the estimation performance even in the median plane. We confirmed that the spectral features in ITF were distinctive in voice frequency band from the measurement of ITF in an anechoic chamber. An experiment was carried out in a real room to investigate the feasibility and the estimation performance of the proposed ear design and the estimation method. In the experiment, we confirmed that both the front-back discrimination and the sound direction estimation can be achieved with reasonable errors. Thus, we expect that robots with the proposed artificial ear can find the direction of user from two output signals only.

ACKNOWLEDGMENT

This research was supported by the National Research Laboratory Program (R0A-2005-000-10112-0), the Brain Korea 21 Project, and the 21st Century Frontier R&D Program.

REFERENCES

[1] C. I. Cheng and G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," *Journal of Audio Engineering Society*, Vol. 49, No. 4, pp. 231-249, 2001.

[2] J. Huang, K. Kume, A. Saji, and M. Nishihashi, "Robotic Spatial Sound Localization and Its 3-D Sound Human Interface," *in proceedings of the first International Symposium on Cyber Worlds*, 2006.

[3] D. Kim, Y. Park, "Explicitly-Adaptive Time Delay Estimation for Wide-Band Signals," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol.E88-A, No.11, p.3214, 2005.

[4] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Tsujino, "Improvement of recognition of simultaneous speech signals using AV integration and scattering theory for humanoid robots," *Speech Communication*, Vol. 44, pp. 97-112, 2004.

[5] H. G. Okuno, K. Nakadai, T. Lourens, and H. Kitano, "Sound and Visual Tracking for Humanoid Robot," *Applied Intelligence*, Vol. 20, No. 3, pp. 253-266, 2004.

[6] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *Journal of Acoustical Society of America*, Vol. 56, No. 6, pp. 1829-1834, 1974.

[7] M. Morimoto and H. Aokata, "Localization cues of sound source in the upper hemisphere," *Journal of Acoustical Society of Japan. (E)*, Vol. 5, pp. 165-173, 1984.

[8] J. Blauert, *Spatial hearing, MIT: The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, MA, 1996.

[9] C. I. Cheng & G. H. Wakefield, "Introduction to Head-Related transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," *Journal of the Audio Engineering Society,* Vol. 49, No. 4, pp.231-248, 2001.

[10] K. Iida, M. Yairi, and M. Morimoto, "Role of Pinna Cavities in Median Plane Localization," *Journal of Acoustical Society of America*, Vol. 103, Issue 5, pp. 2844, 1998.

[11] E. A. G. Shaw and R. Teranishi, "Sound pressure generated in an external-ear replica and real human ears by a nearby point source," *Journal of Acoustical Society of America*, Vol. 44, pp. 240-249, 1968.

[12] R. Teranishi and E. A. G. Shaw, "External-ear acoustic models with simple geometry," *Journal of Acoustical Society of America*, Vol. 44, pp. 257-263, 1968.

[13] E. A. G. Shaw, "The external ear: New knowledge," in Ear Moulds and Associated Problems, *in proceedings of the* s*eventh Danavox Symposium*, Denmark, 1975.

[14] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *Journal of Acoustical Society of America*, Vol. 112, No. 4, pp. 1583-1596, 2000.

[15] E. A. Lopez-Poveda and R. Meddis, "A physical model of sound diffraction and reflections in the human concha," *Journal of Acoustical Society of America*, Vol. 100, No. 5, pp. 3248-3259, 1996.

[16] S. Hwang, K. Shin, and Y. Park, "Artificial ear design for robots," *in proceedings of IEEE SENSORS*, 2006.

[17] J. S. Bendat and A. G. Piersol, *Random Data Analysis and Measurement Procedures*, Wiley, New York, 1999.

[18] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal processing*, Vol. 24, Issue 4, pp. 320-327, 1976.