Robotics and Autonomous Systems 59 (2011) 208–217

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Sound direction estimation using an artificial ear for robots☆

Sungmok Hwang *, Youngjin Park, Youn-sik Park

Marine Research Institute, Samsung Heavy Industries, Co., Ltd., Geoje 656-710, Republic of Korea

ARTICLE INFO

Article history:
Received 2 September 2008
Received in revised form
26 September 2010
Accepted 21 December 2010
Available online 6 January 2011

Keywords:
Sound direction estimation
Artificial ear
Human–robot interaction
Head-related transfer function
Interaural transfer function

ABSTRACT

We propose a novel design of an artificial robot ear for sound direction estimation using two measured outputs only. The spectral features in the interaural transfer functions (ITFs) of the proposed artificial ears are distinctive and move monotonically according to the sound direction. 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 vanish. We propose a localization method to estimate the lateral and vertical angles simultaneously. The lateral angle is estimated using interaural time difference and Woodworth and Schlosberg's formula, and the front–back discrimination is achieved by finding the spectral features in the ITF estimated from two measured outputs. The vertical angle of a sound source in the frontal region is estimated by comparing the spectral features in the estimated ITF with those in the database built in an anechoic chamber. The feasibility of the designed artificial ear and the estimation method were verified in a real environment. In the experiment, it was shown that both the front–back discrimination and the sound direction estimation in the frontal region can be achieved with reasonable accuracy. Thus, we expect that robots with the proposed artificial ear can estimate the direction of speaker from two output signals only.

1. Introduction

Intelligent robots operating in a household environment should detect various sound events and take notice of them to achieve robust recognition and interaction with the user. Thus, robots need to be able to estimate where a sound source is, i.e. 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, and most of them mainly depend on inter-channel time difference (ICTD) or inter-channel level difference (ICLD) with an assumption that the microphones are placed in the free-field or that the head shape is a simple sphere [1–9]. ICTD is defined as the difference in the arrival times of a sound wave-front between two microphones. Likewise, ICLD is defined as the difference in sound pressure levels between two microphones. However, when ICTD and ICLD obtained from two microphones are used, only 1-D estimation is possible because there are many source positions in 3-D space sharing the same ICTD and ICLD, and this is called the "Cone of Confusion" [10]. In other words, the inter-channel differences provide the information about the lateral angle only in the interaural-polar coordinate system[1] as depicted in Fig. 1 (left), and whether a source is in the front or back cannot be distinguished. Thus, many estimation methods based on the inter-channel differences rely on an array of more than two microphones [2–5,9,13,14]. At least 3 microphones are needed to estimate the lateral angle without the front–back confusion [4,13], and more than 4 microphones are needed to estimate both the lateral and vertical angles simultaneously [2,3,14]. In general, the more microphones that are used for sound direction estimation, the more computational load is required to handle microphone signals. Also geometric constraints due to placement of many microphones may conflict with the design of the robot shape. Direction of a sound source also can be estimated by using a vision system, but a source should come within the field of vision and estimation performance can deteriorate under poorly lighted conditions. Recently, many researchers have tried to integrate the audio and visual information for sound direction estimation [1,4,8]. Relatively robust and reliable performance can be achieved by using audio–visual integration, but it leaves much room for improvement because ICTD and ICLD are the only audio information used in the conventional estimation method.

Humans can perceive the sound direction and distinguish whether a sound is coming from the front or the rear, above or

☆ Portions of this work were presented at the International Conference on Control, Automation, and Systems, Oct. 17–20, 2007.

* Corresponding author. Tel.: +82 42 350 3076; fax: +82 42 350 8220.
E-mail addresses: tjdahr78@kaist.ac.kr, sungmok.hwang@samsung.com (S. Hwang), yjpark@kaist.ac.kr (Y. Park), yspark@kaist.ac.kr (Y.-s. Park).

[1] Details on the interaural-polar coordinate system can be found in [11,12].

0921-8890/$ – see front matter
doi:10.1016/j.robot.2010.12.005

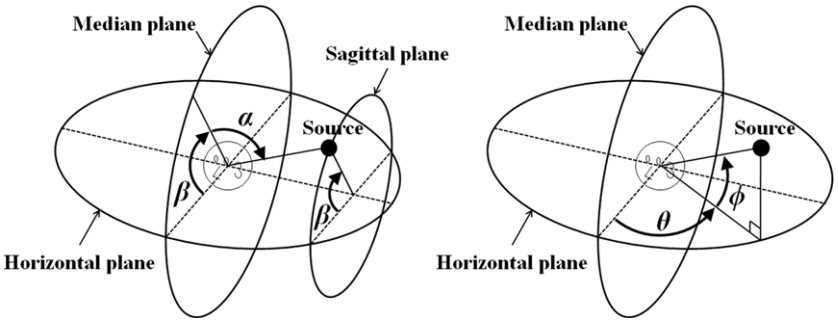

Fig. 1. The interaural–polar coordinate system (left) and the vertical–polar coordinate system (right). α and β are the lateral and vertical angles of a sound source in the interaural–polar coordinate system, respectively. θ and ϕ are the azimuth and elevation angles of a sound source in the vertical-polar coordinate system, respectively.

below with only two ears. It has been discovered by a number of researchers in the past that the human auditory system mainly utilizes primary sound cues, such as interaural time difference (ITD), interaural level difference (ILD), and spectral modification, to perceive sound direction [10,15–20]. ITD and ILD are the same concepts with ICTD and ICLD, respectively. These sound cues are well encrypted in the head-related transfer function (HRTF), which is an acoustical transfer function defined as the ratio of the sound pressure at the eardrum to that measured at the head center with the head absent. The physical structures of a listener, such as head, pinnae, shoulder, and torso, transform the spectrum of sound waves when they reach to the listener's eardrum. This physical transform of sound waves is characterized by the HRTF. It is known that humans mainly depend on the interaural differences, i.e. ITD and ILD, to perceive the lateral angle of a sound source, and utilize the spectral modification to perceive the vertical angle of a sound source [18,21,22]. Especially, spectral modification due to the complex shape of the pinna provides an effective sound cue for the vertical perception [15,19–21,23]. Shaw and Teranishi described that the spectral features including peaks and notches in HRTFs are entailed by the direction dependent acoustic filtering due to the pinna [19,20]. For example, Fig. 2 shows the log–magnitude of HRTF (dB) of a representative subject in the CIPIC HRTF database in the median plane [24]. The abscissa and ordinate indicate the vertical angle in degrees and the frequency (kHz), respectively. The coloured level of a pixel, which represents the magnitude of HRTF in the dB scale, is in accordance with the intensity scale at the right side of the figure. The high frequency characteristics of the HRTFs change substantially with the vertical angle, and the pattern of spectral peaks and notches in the HRTFs is significantly different at each vertical angle. The pattern of spectral features is an effective sound cue for vertical perception. Iida proposed a model for estimation of sound source elevation in the median plane using the first and second notches in the HRTF [25]. The frequencies of elevation-dependent notches were modelled by 4th-order polynomial functions, and the source elevation was estimated by extracting the notches for the ear-input signals and utilizing the polynomial functions.

If an artificial robot ear, which can provide proper sound cues as human pinnae do, is designed properly, it is expected that sound direction in 3-D space can be estimated from two microphone signals only. In other words, robots can utilize the all primary sound cues for sound direction estimation by using the artificial ear, whereas the conventional method depends on ICTD and ICLD only. Thus, the research objective of this study is to design an artificial robot ear by mimicking the human pinna for sound direction estimation in 3-D space based on two microphones only. Detailed design procedure and direction-dependent characteristics of the proposed artificial ear are provided in Section 2. Details on sound direction estimation method using the artificial ear is described in Section 3. To verify the feasibility of the proposed design and angle estimation method, an experiment is carried out in a household environment and the results are included in Section 4.

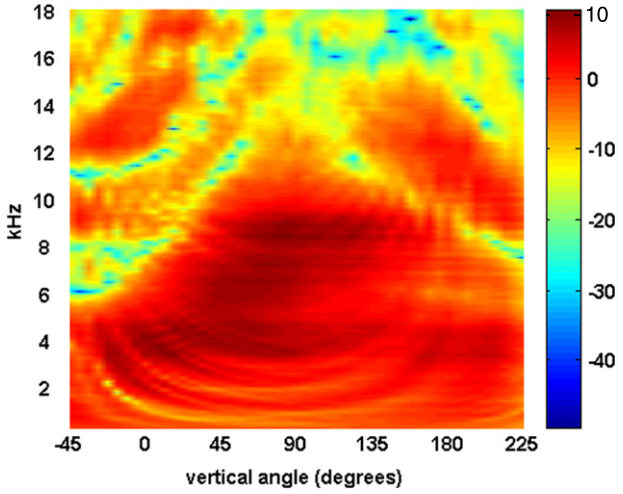

Fig. 2. Log-magnitude of HRTF (dB) of a representative subject in the CIPIC HRTF database in the median plane [24].

2. Artificial ear design

Among the many structures in pinna, it is well known that the posterior wall of the concha is responsible for the spectral notches in HRTFs that serve as important front–back and up–down cues [16,17,23,26]. Hebrank and Wright [16] 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. More recently, Lopez-Poveda and Meddis [24] 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. Inspired by the previous studies, we describe the two kinds of artificial robot ear design mimicking the human concha. Bilateral asymmetric microphone position is also investigated to enhance the estimation performance.

2.1. First design

The first design of an artificial robot ear mimicking the human concha is designed as depicted in Fig. 3. The basic assumption guiding our design is that the angle dependent spectral notch is caused by cancellation between the direct wave reaching the meatus entrance and the reflected wave from the concha posterior wall. With this assumption, we designed the artificial ear so as to