24

Sound Localization of Elevation using Pinnae for Auditory Robots

Tomoko Shimoda, Toru Nakashima, Makoto Kumon,
Ryuichi Kohzawa, Ikuro Mizumoto and Zenta Iwai
Kumamoto University
Japan

1. Introduction

Humans perceive sound lags and differences in loudness between their two ears to ascertain the location of sounds in terms of azimuth, elevation and interval, etc. Humans and animals can determine the direction of sounds using only two ears, by making use of interaural time difference (ITD), interaural phase difference (IPD), interaural intensity difference (IID), etc. Robots can also use sound localization to detect changes in the environment around them; consequently, various sound localization methods for robots have been investigated. Since many animals have two ears, two ears seem to be the minimum requirement for determining sound localization.

While most robot's sound localization systems are based on microphone arrays that consist of three or more microphones, several researchers have attempted achieving binaural sound localization in robots. Nakashima (Nakashima and Mukai, 2004) developed a system that consisted of two microphones with pinnae and a camera that was only utilized in learning. By adopting a neural network that estimated the control command, he proposed a method utilizing sound signals for guiding the robot in the direction of a sound. Takanishi (Takanishi et al., 1993) achieved continuous direction localization of a sound source in the horizontal plane. They used the interaural sound pressure difference and the ONSET time difference as parameters and proposed a method for achieving two-step direction localization in a humanoid head using these parameters. Kumon (Kumon et al., 2003) approached the problem using an adaptive audio servo system based on IID as the control method for orientating the robots in the direction of the sound source in the horizontal plane by using two microphones.

This chapter describes the use of an audio servo of elevation for achieving sound localization using spectral cues caused by pinnae. Here, the term "audio servo" denotes a method for simultaneously localizing sound sources and controlling the robot's configuration, in contrast with conventional methods that are based on the "listen-and-move" principal. The audio servo is characterized by a feedback system that steers the robot toward the direction of the sound by combining dynamic motion of the robot with auditory ability. In order to achieve this, this chapter proposes a method for detecting instantaneous spectral cues, when these cues are not very accurate. Furthermore, controllers that compensate for the inaccuracy of the measured signal are also considered. In addition, this

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel, ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

chapter considers using sound source separation to detect spectral cues even in a noisy environment by attenuating the noise. Sound source separation is achieved using two microphones while spectral cues from only a single microphone and a pinna are utilized.
This chapter is organized as follows. In the next section (Section II) spectral cues caused by pinnae and artificial pinnae are briefly introduced. Then, in Section III a robust detection method for spectral cues is described. Methods for measuring spectral cues and a filter for investigating the relationships between the elevation angle and spectral cues are described. Modelling and identification with respect to the relationship between the elevation angle and the filtered spectral cues are also presented in this section. This section also describes sound source separation. Section IV describes the realization of an audio servo system that includes a controller for an auditory robot; its performance and results from experiments are also presented. Finally, Section V gives the conclusions and describes future projects.

2. Spectral Cues

This section briefly introduces spectral cues, pinnae and their frequency responses.

2.1 Spectral Cues

Generally, humans are considered to use frequency domain cues to estimate the elevation of a sound source (Garas, 2000). The frequency response varies with respect to the sound source direction as a result of the interference that occurs between the sound wave that enters the auditory canal directly and the sound wave reflected from the pinnae. In particular, spectral peaks and notches produced respectively by constructive and destructive interference contain information regarding the elevation of the sound source, making it possible to estimate the elevation of a sound source by analyzing them.

2.2 Robotic Pinnae

Spectral cues are dependent on the shape of the pinnae. In this chapter, logarithmic-shaped reflectors were used as pinnae (see Fig. 1). The pinnae had a depth of 6 (cm) (Lopez-Poveda and Meddis, 1996) and were made from 0.5 (mm) thick aluminum sheets. Figure 2 shows a photograph of experimental device with the pinnae attached. Figures 3 and 4 show a front view and a side view of the experimental device with the pinnae attached, respectively.

Figure 1. Developed Pinnae

Figure 2. Photograph of robot with pinnae attached

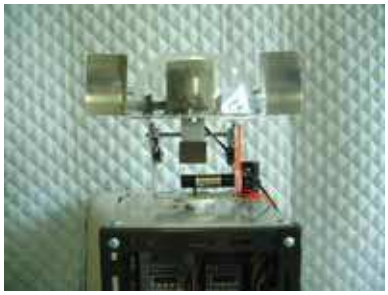
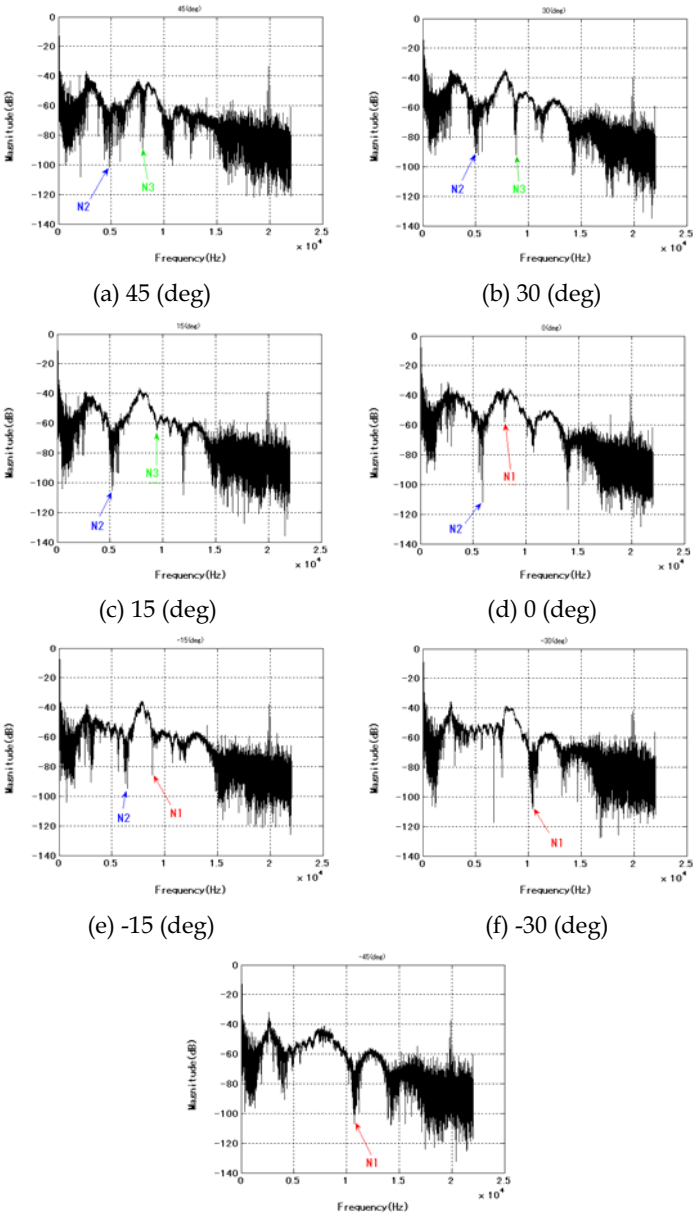


Figure 3. Front view of the robot with pinnae

Figure 4. Side view of the robot with pinnae

The frequency response of the developed robotic pinna was measured to examine the relationship between spectral cues and sound source elevation. The robot's head was kept still while these measurements were made. A loudspeaker was positioned 0.5 (m) in front of the robot. The frequency characteristics of the pinnae were measured using time-stretched pulses (TSPs). The sound source direction is expressed as follows. The angle is defined as being 0 (deg) when the sound source is located directly in front of the robot. When the sound source is located below the robot's head, the angle is denoted by a positive value.

The results obtained using TSP are shown in Figs. 5(a) to (g). In these results, there are three sharp notches (labeled N1, N2 and N3) within the frequency range from 2 (kHz) to 15 (kHz) and these notches shift to lower frequencies as the robot turned its head upward. Thus, it can be concluded that it is possible to detect the elevation angle of a sound source using pinna cues.



(a) 45 (deg) (b) 30 (deg)

(c) 15 (deg) (d) 0 (deg)

(e) -15 (deg) (f) -30 (deg)

(g) -45 (deg)

Figure 5. Frequency responses of the developed pinnae

3. Spectral Cues Detection

3.1 Spectral Cues Extraction

In what follows, spectral cue candidates are defined as the center points between adjacent peaks and notches in a smoothed power spectrum; these candidates are simply referred to as spectral cues in this chapter. The following method is adopted to determine the frequency of a spectral cue. Firstly, the sound signal is transformed using short-time fast Fourier transformation (STFFT) and the frequency response is obtained for each instance of time (Fig. 5). Next, this frequency response is smoothed using a zero-phase low-pass filter (LPF).

An example of dynamic spectral cues is shown in Fig. 6. The ordinate axis represents frequency (Hz) and the abscissas axis represents time (s). Each of points represents a spectral cue. The sound source was fixed 0.5 (m) in front of the robot, as above, while the head was programmed to follow a triangular reference trajectory (Fig. 7). A white signal was utilized as the pilot signal instead of a TSP. Although multiple spectral cue candidates exist at every instant in Fig. 6, most of them varied with the head motion.

Unfortunately, not all of the detected candidates corresponded with the head motion, and there are some frequency bands in which spectral cues disappear for a specific direction. In addition, several adjacent cues had similar frequencies as each other. Hence it is difficult to immediately determine the sound source direction for these candidates. In particular, since sound signals in a short sampling time were employed, it is possible that the time frequency responses could have been degraded due to extraneous noises. Thus an improved robust cue detection method is required; such a method is described in following subsection.

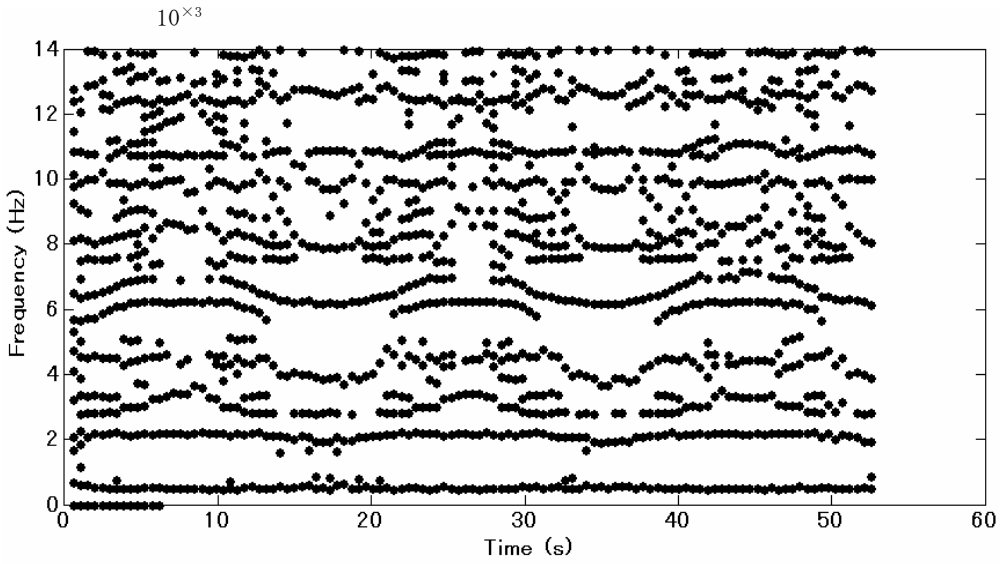


Figure 6. Spectral Cue Candidates obtained by experiment

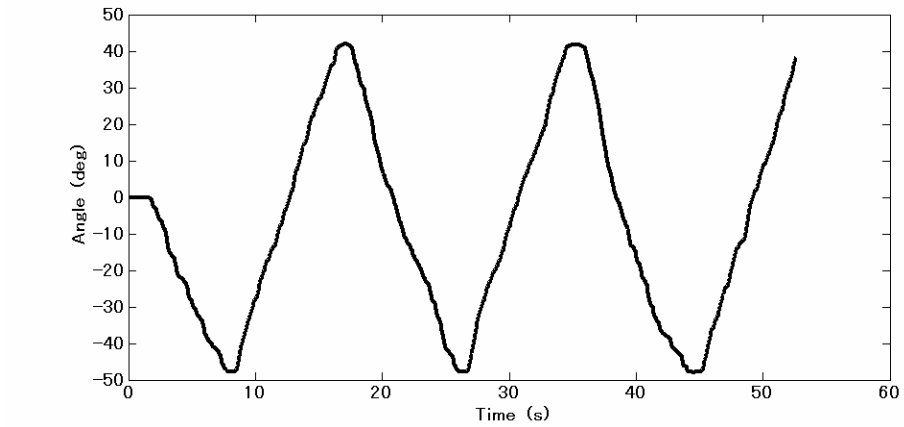


Figure 7. Head motion of the robot for the experiment in which the spectral cues of Fig. 6 were obtained

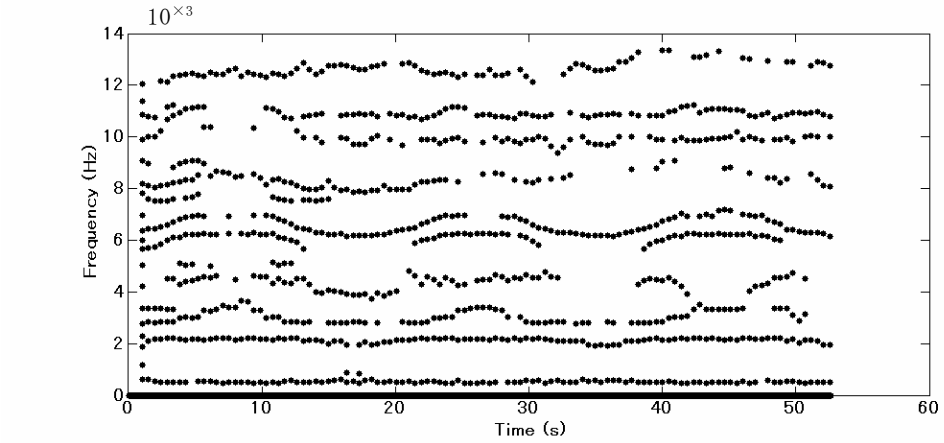


Figure 8. Clustered spectral cues produced by the proposed method (corresponds to Fig. 6)

3.2 Clustering

The robot's head motion is continuous owing to its dynamics. Since there is a relationship between the spectral cues and sound source direction, it is conceivable that the variation of spectral cues due to the motion of the robot's head is also continuous. Thus, given the values of the spectral cues at any particular moment, it should be possible to determine the spectral cues for the next time step by searching in their neighborhoods. Based on this supposition, a filter for detecting spectral cues from the frequency response for each instant was designed as follows.

Let $\boldsymbol{f}_k$ denote the vector whose elements represent the spectral cues' frequencies at time k; f_{kn} represents the n-th element of this vector. We assume that the frequency of a spectral cue remains within a certain range, which is referred to as the scope. The validity of the

measured signal is stored in a vector of flags whose i-th element is denoted as d_{ki}. We assume that the number of cues is given and we denoted it by N_k. The proposed algorithm is given as follows.

<Spectral Cue Clustering>

1. Given spectral cues at time k.
2. Measure the frequency response at time $k+1$. Obtain the candidates of the spectral cues and denote them by the vector $\hat{f}_{k+1}$ whose n-th element ($\hat{f}_{(k+1)n}$) represents the frequency of the candidate.
3. Compute the assessment function $J_n(r)$ as follows.

$$J_n(r) = \begin{cases} C\Delta(n,r) & (\Delta(n,r) \geq 0), \\ -C\Delta(n,r) & (\Delta(n,r) < 0), \end{cases} \tag{1}$$

where $\Delta(n,r) = f_{kn} - \hat{f}_{(k+1)r}$ and $0 < C < 1$.

For each n, find r that minimizes $J_n(r)$ and let $f_{(k+1)n} = \hat{f}_{(k+1)r}$ as far as $\hat{f}_{(k+1)r}$ lies within the r scope. Otherwise, let $f_{(k+1)n}$ be f_{kn}.

4. If $f_{(k+1)n} = f_{(k+1)(n+1)}$, then replace $f_{(k+1)(n+1)}$ with $f_{(k+1)(n+1)}+\delta$, where δ represents a small positive constant.
5. If $\hat{f}_{(k+1)r}$ in step 3 lies outside of the scope, let the flag $d_{(k+1)i} = 0$. Otherwise let $d_{ki} = 1$. Note that d_{ki} will be also updated using the sound separation method described below.
6. Return to step 2 by incrementing k by 1.

Figure 8 shows the clustered spectral cues of Fig. 6 processed using the above method. The initial frequencies of cues were defined every 300 (Hz) from 0 (Hz) to 6000 (Hz). Scopes were given as ranges of 300 (Hz) around their initial frequencies.

3.3 Modeling

In the previous section, an algorithm for detecting spectral cues was proposed. Next, the relationship between the filtered spectral cues and the sound source detection is considered in order to localize the sound source direction correctly. The simplest model that determines the elevation angle from the frequency, the relationship between the filtered spectral cues and the sound source direction is introduced. Let θ_e be the angular difference between the sound source and the robot's head. Let the frequency of the filtered spectral cues and coefficients that are identified below be denoted by f_i, C_i and C_{i0}, respectively. The model is then expressed mathematically by:

$$\theta_e = \frac{\sum_i \left(C_i d_i f_i + C_{i0} d_i\right)}{\sum_i d_i}, \tag{2}$$

where d_i is 1 or 0.

3.4 Identification

Just as when measuring the characteristic of the pinnae, the sound source was fixed approximately 0.5 (m) in front of the robot. The white signal was generated while the robotic head was in motion. The data measured when d_i was not 0 were used to determine C_i and

C_{i0}. Using these extracted data, C_i and C_{i0} are determined such that they minimize the following squared residual,

$$\sum_{k \in K_i} (\theta(k) - (C_i f_i(k) + C_{i0}))^2, \tag{3}$$

where K_i represents a set of times when $d_i \neq 0$ and $\theta(k)$ defines the direction of the sound source computed using the angle of the robot's head data.

The elevation angles estimated by the above method with the identified coefficients are shown in Fig. 9. The ordinate axis represents the angle of the robot's head while the abscissas axis represents time. The solid line indicates the motion of the robot's head while the points indicate the computed angles; there is good agreement between the two. Figure 9 was produced using three spectral cues that correspond to the motion of the robot's head; this figure confirms the method used for estimating the elevation angle of the sound source.

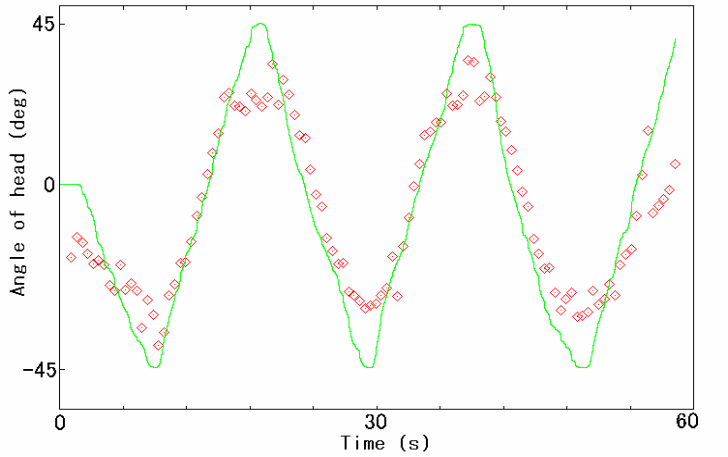


Figure 9. Elevation angle estimated by the model expressed by Eq. (2)

3.5 Sound Source Separation

In the previous sections, sound source localization was considered in order to adapt the method described above to cases when there is a lot of noise present. Accurate spectral information is critical for performing sound localization with spectral cues. In real applications, however, spectral information is often contaminated with extraneous noise. It is thus necessary to separate the extraneous noise from the target noise and to extract only the spectral cues from the target sound source.

In order to achieve this, horizontal sound source separation with two microphones is performed by assuming that the sound source and the extraneous noise do not originate from the same median plane. Although this assumption is rather strict, it is acceptable for many practical situations. In order to separate horizontal sound sources, Suzuki's method (Suzuki et al., 2005) was adopted; this method focuses on the proportional relationship that exists between the IPD and frequency.

3.5.1 Horizontal sound source separation (Suzuki et al., 2005)

Consider two microphones installed with a displacement a. When a planar sound wave with frequency f propagates in the direction ξ (see Fig. 10), the phase difference $\Delta\varphi(f)$ is given by

$$\Delta\phi(f) = \frac{a \sin\xi}{V} f, \tag{4}$$

where V represents the speed of sound. Therefore, the relationship between f and φ can be expressed by

$$f = \alpha\Delta\phi, \tag{5}$$

where α is given by $V/a \sin\xi$ and is treated as a constant. Equation (5) for the case when there is only one sound source is depicted in Fig. 11. The ordinate and abscissas axes represent the frequency and the phase difference $(\Delta\phi, f)$, respectively. Each data point belongs to a line passing through the origin. Thus line detection can be utilized for separating sound sources.

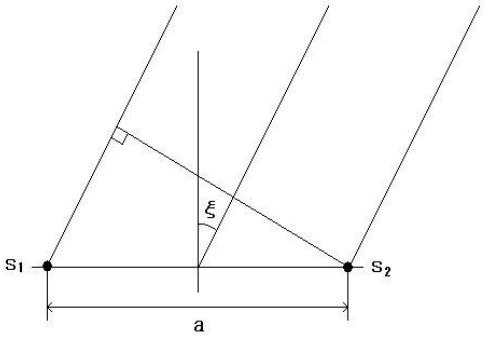


Figure 10. Diagram showing measurement of IPD using two microphones

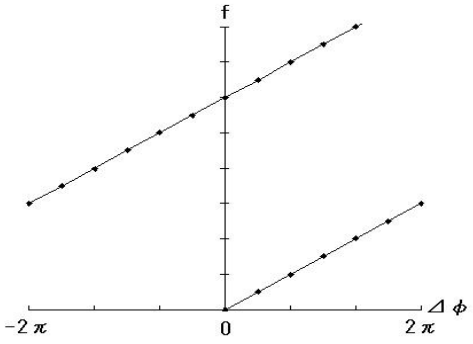


Figure 11. Signal from a single sound source depicted in frequency-IPD space

3.5.2 Line Detection

In line detection, the gradient ($\alpha = f / \Delta\phi$) is first computed at each measured point $(\Delta\phi, f)$. The angle (η_0) of the straight line connecting $(\Delta\phi, f)$ is defined as follows.

$$\eta_0 = \tan^{-1} \frac{f}{\Delta\phi}. \tag{6}$$

Since the phase difference is circular periodic, shifted values of α, $f/(\Delta\phi + 2\pi)$ and $f/(\Delta\phi - 2\pi)$ are also taken into consideration as shown in Fig. 12.

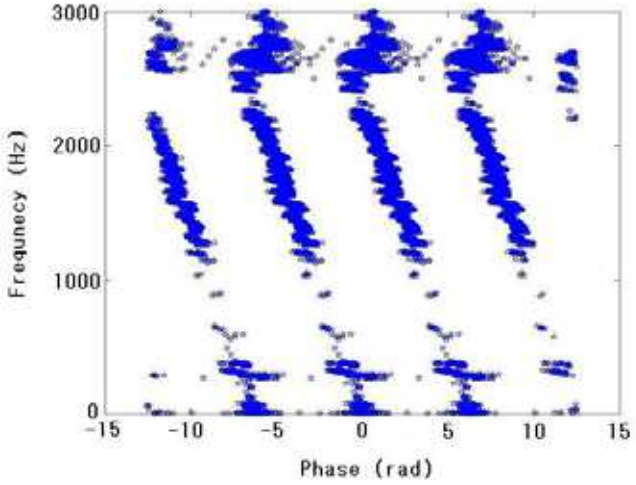


Figure 12. Measured response in frequency and IPD

In order to determine η_0, let us consider the relationship between *f* and $\Delta\phi$ of the measured data (Fig. 12). These data are plotted in two dimensions, where the ordinate axis represents frequency (*f*) and the abscissa axis represents phase difference ($\Delta\phi$). The space is discretized by dividing it into *2n* fan-shaped regions whose vertical angles at the origin are given by $(\pi - 2\eta_{\min})/2n$. $\eta_{\min}$ is the vertical angle when the sound source is located just beside the robot, or when the phase difference is minimized.

The line of the sound source is computed by determining the region to which it belongs. This is done by counting the numbers of $(\Delta\phi, f)$ points that exist in each region. Region *j* is defined by the angle η_0, which satisfies the condition

$$\eta_{j-1} \le \eta_0 < \eta_j, \tag{7}$$

where

$$\eta_j = \eta_{\min} + j\frac{\pi - 2\eta_{\min}}{2n}.$$

Let *P(j)* represent the number of points in the region *j*. Given a point (ξ), *j* is expressed as follows.

$$j = \left\lceil \frac{2n}{\pi - 2\xi_{\min}} (\xi_0 - \xi_{\min}) \right\rceil + 1, \tag{9}$$

where $\lceil X \rceil$ represents the largest integer that is smaller than *X*. By using the above expressions, *P(j)* is counted for all data points. For simplicity, in this chapter, the region having the most points is taken to be the region of the sound source. Alternative selection algorithms are also possible and they should be investigated in the future.
Consequently, the region that contains spectral cues from the sound source can be determined by evaluating the frequency of points in the region having the most data points. If this region does not contain the above-mentioned detected spectral cue, then d_i is set to 0. Spectral cues from the sound source are extracted by determining d_i.

4. Audio Servo

As mentioned in the introduction, audio servo is a method for simultaneously achieving sound localization and configuration control. In this section, measured spectral cues are utilized in an audio servo system. The robotic controller is derived first and its performance is then evaluated. Finally, an experiment involving this controller and the results obtained are described.

4.1 Controller Design

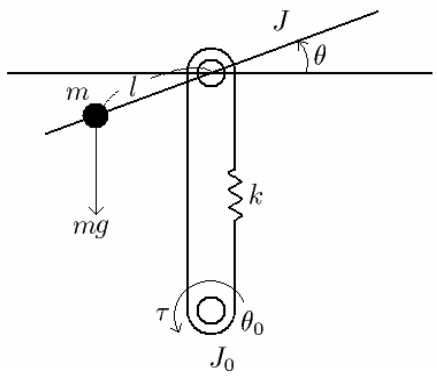


Figure 13. Dynamical model of the robot head.

Figure 13 shows the model for the robotic head; the equations of motion for this model are given as follows:

$$\begin{cases} J\ddot{\theta} = mgl\cos\theta - ksr - d\dot{\theta}, \\ J_0\ddot{\theta}_0 = ksr + \tau, \end{cases} \tag{10}$$

where J is the inertia of the upper pulley, J_0 is the inertia of the lower pulley, k is the elastic constant of the belt, m is the mass of the robot's head, g is the acceleration of gravity, l is the distance from the center of gravity, d is the coefficient of friction factor, and θ and θ_0 are the rotational angles of the robot's head and the motor shaft, respectively. s is given by $r\theta - r_0\theta_0$, where r and r_0 are the radii of the upper and lower pulleys, respectively.
The motor torque $\tau$is modeled by

$$\tau = -K(\dot{\theta}_0 - u), \tag{11}$$

where K is the feedback gain of the servo system and u is the control input which is defined below.
The elevation angle of the sound source is denoted by θ_d and define $e = \theta - \theta_d$. The output is defined by $y = \hat{\theta} + \gamma\dot{\theta}_0$ and the state of the system as $\mathbf{x} = (\theta - \theta_d \quad \theta^{(1)} \quad \theta^{(2)} \quad \theta^{(3)})^T$. γ is an arbitrary positive value and it is given as a small constant below since its exact value is not critical. Assuming that $|\psi|\mathcal{E}|^1$ is small, the dynamic model of the robot can be approximated as:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u + \Delta, \\ y = \mathbf{C}\mathbf{x}, \end{cases} \tag{12}$$

where $\mathbf{A} \in \mathbf{R}^{4\times4}$, $\mathbf{b} \in \mathbf{R}^4$ and $c \in \mathbf{R}^{4\times1}$ represent matrices. Δ is a vector of nonlinear functions. Specifically,

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & a_1 & a_2 & a_3 \end{pmatrix}, b = \begin{pmatrix} 0 \\ 0 \\ 0 \\ a_6 \end{pmatrix}, c = (c_1 \quad \frac{r}{r_0}\gamma \quad \frac{d}{krr_0}\gamma \quad \frac{J}{krr_0}\gamma), \ \Delta = \begin{pmatrix} 0 \\ 0 \\ 0 \\ a_4\cos e + a_5 \sin e \end{pmatrix}$$

where

$$a_1 = -(\frac{Kkr^2 + krr_0 d}{JJ_0}), \ a_2 = -\frac{kr(Jr_0 + J_0 r + Kd)}{JJ_0}, \ a_3 = -\frac{d + KJ}{JJ_0},$$

$$a_4 = \frac{krr_0 mgl}{JJ_0}\cos\theta_d, \ a_5 = -\frac{krr_0 mgl}{JJ_0}\sin\theta_d, \ a_6 = \frac{Kkrr_0}{JJ_0}.$$

The dominant system (***A, b, c***) of (12) satisfies the almost strictly positive real (ASPR) condition which implies that it is possible to achieve high gain output feedback using the control objective (Wen, 1988, Kaufman, 1998). Hence, the control input u is given as,

$$u = -My, \tag{13}$$

where M is an appropriate positive constant. It should therefore be possible to align the robot with the sound source direction (Kumon et al., 2005).
Figure 14 shows the result of the experiment when sound separation was used. The ordinate axis represents angle from the horizontal plane (deg) and the abscissas axis represents time (s). The target sound source was located 16 (deg) above the horizontal plane. After about 10

(s), the robot was aligned in the direction of the sound source, although it oscillated about the true direction.
By way of comparison, Fig. 15 shows the result of the experiment when sound separation was not employed. It demonstrates that it is necessary to determine the frequency domain which contains the spectral cues of the sound source.

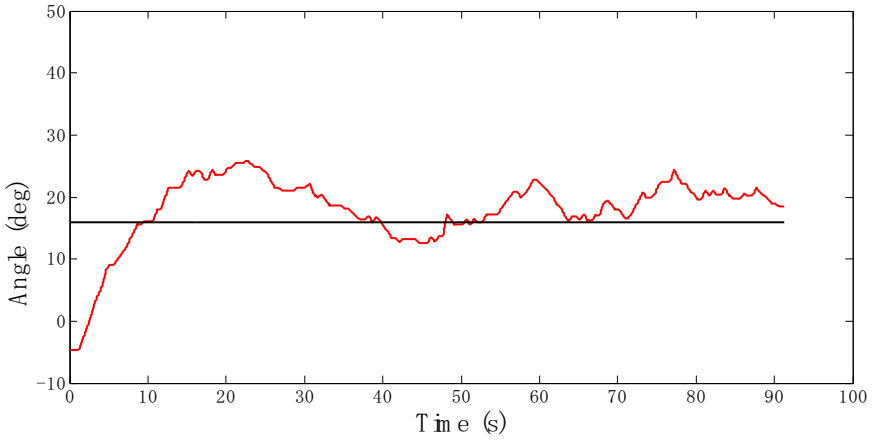


Figure 14. Motion of the robot's head with sound separation

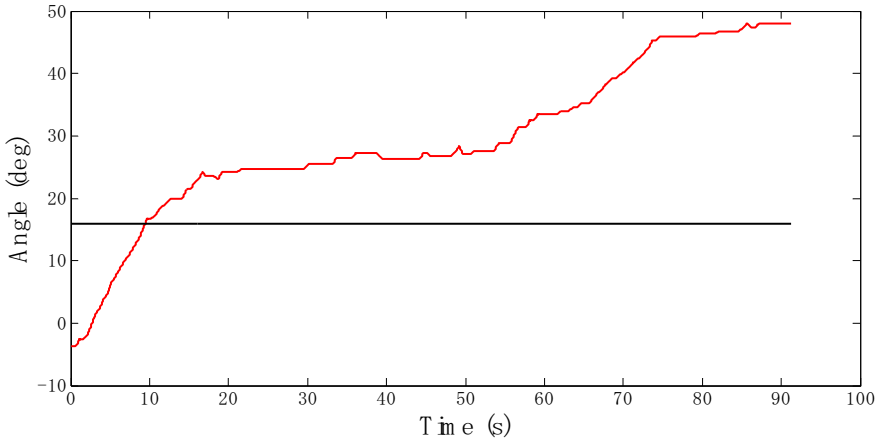


Figure 15. Motion of the robot's head without sound separation

4.2 Controller performance

In the above experiment, the robot head oscillated about the correct orientation; this aspect needs to be improved. When controlling robots based on the location of sound sources, it is clearly desirable to precisely locate cues and to accurately determine the relationship between cues and physical quantity. However, it is impractical to rely on sufficient measurement accuracy when the number of microphones is restricted and when external noise is present. We therefore evaluated the performance of the audio servo controller by

accounting for inaccurate measurements. The performance of the audio servo controller was evaluated after the structure of the controller had been modified (13) in order to attenuate the effect of noise.

4.2.1 Controller

When the effect of observation noise is considered, the system is modeled using the following set of equations rather than that of (12),

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u + \Delta, \\ y = \mathbf{C}\mathbf{x} + h, \end{cases} \tag{14}$$

where h is the observation noise and it is assumed to be bounded. When observation noise h is present, using a higher feedback gain M in (13) does not necessarily improve control performance, since it also increases the sensitivity to the observation noise h.

The dead zone is one control technique that can be applied to noisy output signals; it disregards the output, or error, if magnitude of the output is smaller than a given threshold. The following modified controllers were evaluated by applying the dead zone:

$$u = \begin{cases} -My & (|\, y_0 \,| < y) \\ 0 & (|\, y_0 \,| > y) \end{cases} \tag{15}$$

$$u = \begin{cases} -M(y - y_0) & (y > y_0) \\ 0 & (|\, y_0 \,| > y) \\ -M(y + y_0) & (y < -y_0) \end{cases} \tag{16}$$

$$u = \begin{cases} L & (y > y_0) \\ 0 & (|\, y_0 \,| > y) \\ -L & (y < -y_0) \end{cases} \tag{17}$$

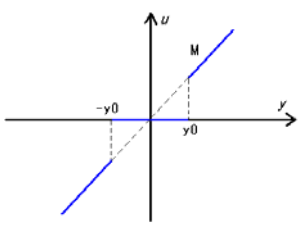


Figure 16. Controller (15)

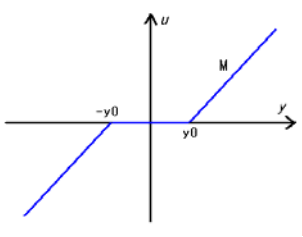


Figure 17. Controller (16)

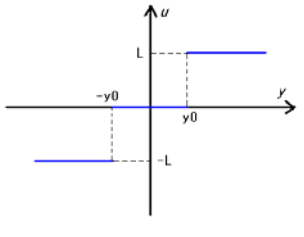


Figure 18. Controller (17)

4.2.2 Experiment and Performance

The proposed controllers were implemented in the robot and their performances were evaluated. In the experiment, the sound source was positioned in front of the robot. The initial elevation angle of the sound source was approximately 16 (deg) above the horizontal plane. A white signal was generated for about 3 (min), which included the duration of the experiment. Three different gains were used for each of the three controllers, namely M=0.1, 0.5, 1.0 for (15) and (16), and L=10, 50, 100 for (17).

Figures 19, 20 and 21 show the motion of the robot's head in these experiments when the parameters of the controllers were M=0.5, L=50 and y_0=3 (deg), respectively. In all these figures, the ordinate axis is the angle of the robot's head (deg) and the abscissa axe is time (s).

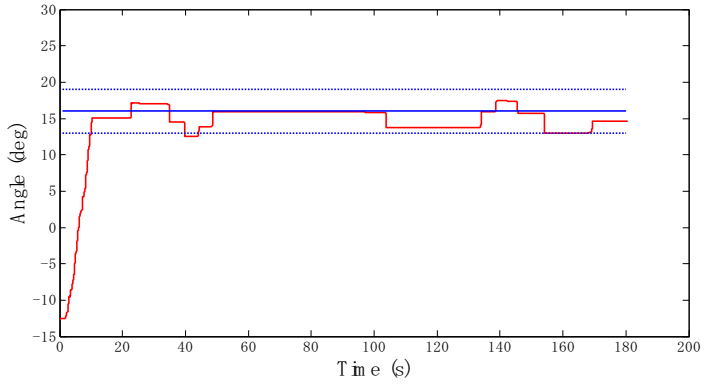


Figure 19. Result with Controller (15)

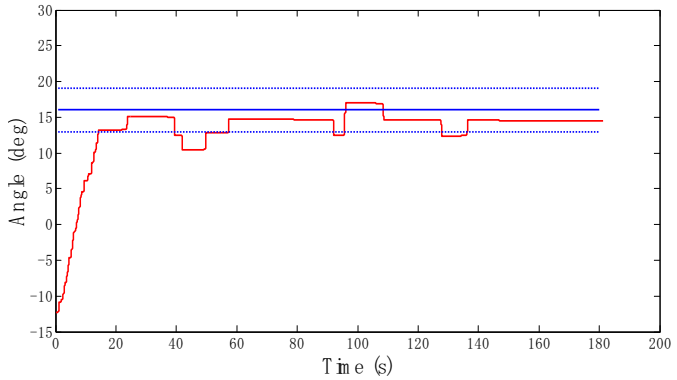


Figure 20. Result with Controller (16)

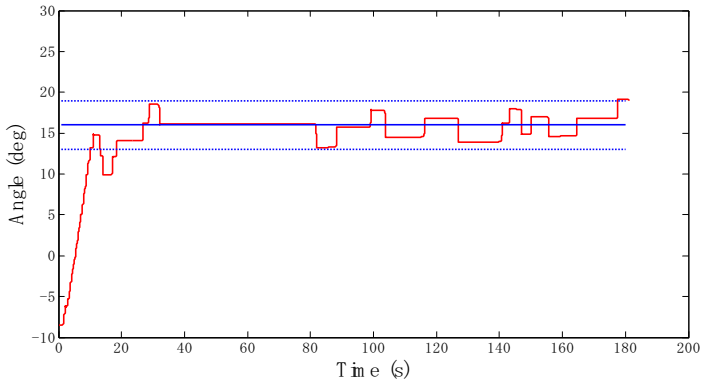


Figure 21. Result with Controller (17)

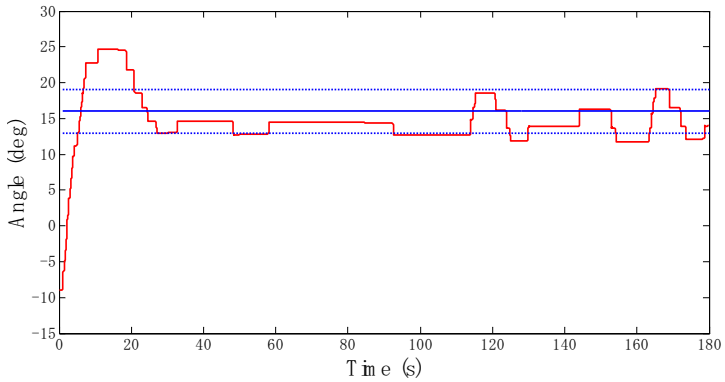


Figure 22. Result of Controller (16) for high gain (M=1.0)

Controller	Gain	Transient Period (s)	Means Error (deg)	Standard Deviation (deg)
(13) with sound separation	1.0	8.6	1.6	57.4
(15)	1.0	9.9	4.5	2.2
(15)	0.5	10.1	0.7	1.2
(15)	0.1	47.8	3.3	1.1
(16)	1.0	5.4	1.0	3.1
(16)	0.5	14.1	1.8	1.5
(16)	0.1	78.0	0.8	0.8
(17)	100	7.2	1.8	2.3
(17)	50	10.0	0.5	1.6
(17)	10	50.0	2.3	2.2

Table 1. Transient Period Error Variance

All of the results show that the robot's head was stably oriented toward the sound source. However, as the gains M, L increase, oscillations were observed with all the controllers. For instance, Fig. 22 shows the case when gain M is 1.0. The mean error angle and variance after the robot's head had been stably oriented toward the target are also shown in this figure.

Table 1 shows the transient period (s), mean error (deg) and standard deviation (deg) for each combination of controller and gain used. When the robot's head was oriented towards the sound source, the estimated angle had a standard deviation 2.1 (deg).

From Table 1, it can be concluded that all controllers succeeded in controlling the robot's head so that it was oriented toward the sound source and that the dead zone technique was effective in attenuating the vibration. However, if a high gain is used to achieve faster response, the control performance deteriorates since the robot moves in an oscillating manner. On the other hand, if a *gray* gain is used, it is possible to achieve a better response

speed and convergence performance. In particular, it is noteworthy that the deviation achieved is better than that for the uncontrolled case.

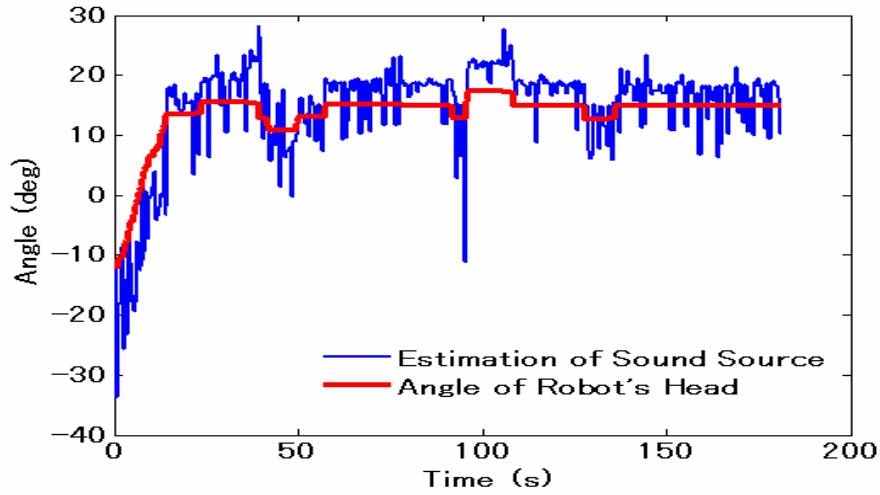


Figure 23. Estimation of Sound Source and Angle of Robot's Head

Figure 23 shows the comparison of the estimated sound source direction and the angle of robot's head. The blue line shows the estimated angle of the sound source and it has a standard deviation of 5.0 (deg), while the red line shows the angle of robot's head and it has a standard deviation of 1.5 (deg). Thus the performance of the sound source localization is improved if the angle of the robot's head is utilized rather than the direct estimated signal, confirming the effectiveness of the technique.

In summary, even when the robot's auditory sensing is inaccurate to some extent, it has been demonstrated that the performance of audio servo can be improved by using it in combination with an appropriately designed control system. If the use of an actual robot head is possible, this technique could be used to improve its sound localization ability by using an audio servo.

5. Conclusion

In this chapter, by using a system consisting of two microphones and one pinna, a method for sound localization using spectral cues was considered. In particular, a robust spectral cue detection method was considered and a method for orientating the robot's head toward a sound source was proposed. In addition, this chapter considers the use of sound source separation in order to attenuate the effect of noise. The conclusions of this present study are summarized as follows:

- Real robotic pinnae were designed and a robot using the pinnae was developed.
- In order to realize sound localization with vertical displacement, an algorithm for detecting spectral cues using the developed pinnae was proposed.
- Spectral cue detection was made robust by considering their frequency continuity with respect to time. A model for determining the sound elevation angle by measuring spectral cues was introduced.

- Horizontal sound separation was considered for extracting the sounds of interest in a noisy environment so that the robot would recognize the spectral cues of only the sound source.
- Controllers were designed to realize an audio servo using the derived spectral cues by the proposed algorithm.
- The performances of the audio servo controllers were evaluated by accounting for the need to allow measurement error when designing the controllers.
- Experiments were conducted with the developed robot. The results demonstrated the ability of the system to orientate the robot's head in the direction of the sound source.

While the proposed method was able to localize the sound source vertically, higher precision sound localization is needed. In order to achieve this, we intend to extend our research as follows:

- Online identification of coefficients C_i and C_{i0} is required for practical applications.
- Determination of optimal gains and parameters is also required.
- In this chapter, a white signal was used as the target. However, sound localization of other sound sources, such as a human voice, is also needed.
- In this chapter, sound localization in only the vertical direction is described. In future work, sound localization should also include lateral and distance detection, thus making it possible to work in three-dimensional space.

6. References

J.Garas, (2000). Adaptive 3D Sound Systems, Kluwer.

H.Kaufman, I.Bar-Kana and K.Sobel, (1998) Direct Adaptive Control Algorithms: Theory and Application.

M.Kumon, T.Sugawara, K.Miike, I.Mizumoto, and Z.Iwai, (2003). Adaptive audio servo for multirate robot systems, *Proceeding of 2003 International Conference on Intelligent Robot Systems,* pp.182-187.

M.Kumon, T.Shimoda, R.Kohzawa, I.Mizumoto, and Z.Iwai, (2005). Audio Servo for Robotic Systems with Pinnae, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.885-890.

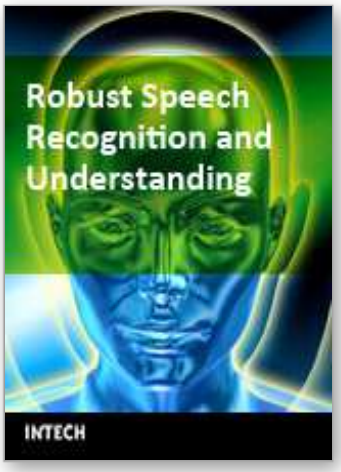
E.A.Lopez-Poveda and R.Meddis, (1996). A physical model of sound diffraction and reflections in the human concha, The Journal of the Acoustical Society of America, vol.100, no.5, pp.3248-3259.

H.Nakashima, T.Mukai, (2004). The Sound Source Localization Learning System for a Front Sound Source, *22nd Annual Conference of the Robotics Society of Japan*.

K.Suzuki, T.Koga, J.Hirokawa, H.Ogawa, and N.Matsuhira, (2005). Clustering of sound-source signals using Hough transformation, and application to omni-directional acoustic sense for robots, Special Interest on AI Challenges Japanese Society for Artificial Intelligence, pp.53-59 (In Japanese).

A.Takanishi, S.Masukawa, Y.Mori and T.Ogawa, (1993). Study on Anthroporomorphic Auditory Robot ~Continuous Localization of a Sound Source in Horizontal Plane, *11th Annual Conference of the Robotics Society of Japan*, pp.793-796.

H.T.Wen, (1988). Time Domain and Frequency Domain Conditions for Strict Positive Realness, *IEEE Transaction on Automatic Control*, 33-10, pp.988-992.



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821