A LOCALIZATION METHOD FOR MULTIPLE SOUND SOURCES BY USING COHERENCE FUNCTION

Hiromichi NAKASHIMA, Mitsuru KAWAMOTO, and Toshiharu MUKAI

RIKEN RTC Research Center
Anagahora, Shimoshidami, Moriyama-ku, Nagoya 463-0003, Japan
email: nakas@nagoya.riken.jp, tosh@nagoya.riken.jp

Advanced Industrial Science and Technology
2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan
email: m.kawamoto@aist.go.jp

ABSTRACT

In the present study, we propose a technique for estimating the vertical and horizontal directions of sound sources using a robot head incorporating two microphones. The microphones have reflectors that work as human's pinnae. We previously proposed a localization method that uses the reflectors to estimate the 3D direction of a single sound source. The present paper deals with the problem of 3D localization of multiple sound sources. Since silent intervals normally occur when a person speaks, such intervals are thought to convey that only one person is talking, even if two or more people are actually speaking. These intervals can be estimated by calculating the correlativity of two audio signals obtained from two microphones. The correlativity is determined by a coherence function. Using the interval for each sound source, we can detect the 3D directions of multiple sound sources. Experimental results demonstrate the validity of the proposed method.

1. INTRODUCTION

When a robot and person talk with each other, it is desirable for the robot to be facing the speaker. Therefore, the robot must be able to detect the direction of the speaker. Moreover, when two or more sounds exist, the sounds detected by the microphones of the robot are mixed, which is typical situation for real environments in which the robot may actually be used. In the environment, the robot hears the mixed sounds and must detect the direction of each sound source from the observed sounds.

Many localization methods have been proposed until now. For example, the model of Ono et al. [1] can be located to one sound source that exists in an upper and lower, right and left direction, using two microphones. This model mimics the function of the ear of the Barn owl, and uses two directional microphones arranged in right and left asymmetry. In case of symmetrically arranged microphones, the Interaural Time Difference (ITD) and the Interaural Level Difference (ILD) are 0 for a sound in the median plane. In the case of asymmetrically arranged microphones, since the ITD and the ILD are not always 0 for a sound in the median plane, the direction of the sound source can be determined by combining the ITD and the ILD.

In the present paper, we propose a technique for estimating the vertical and horizontal directions of sounds using a robot head that incorporates two microphones which have reflectors that work like pinnae. The original point of the proposed localization technique is that the single sound source of the voice is obtained by calculating the correlativity of two audio signals detected by the two microphones, where the observed audio signal is such a signal that two or more sounds are mixed. The correlativity is determined by a coherence function. Using the detected single sound source, its horizontal and vertical direction is estimated.

2. PROPOSED METHOD FOR LOCALIZING TWO OR MORE SOUND SOURCES

Many silent intervals exist when a person is speaking normally. Hence, even if two or more people are speaking, it occurs such a interval that only one person is talking. In the present study, the interval is used to estimate each direction for multiple sound sources. In order to detect the single sound source part, a technique using the coherence function proposed by Mohan is used [2]. Then, our proposed method [3] is used in order to estimate each direction. The flow of the actual processing is as follows:

1. Calculate the value of the coherence function of the voice detected with the two microphones.
2. Based on the processing results, find the interval in which there is only one sound source and extract that particular voice.
3. Process the sound localization using the extracted audio signal.

2.1 Detection of a single sound source

The method of finding the section of one sound source is described first. The input signals from two microphones are assumed to be $y_1(n)$ and $y_2(n)$. These input signals are partitioned into overlapping N-sample time-blocks. $Y_1(m,\omega_k)$ and $Y_2(m,\omega_k)$ are assumed to be the discrete Fourier transforms by the Nth points of $\{y_1(n)\}_{n=mK}^{mK+N-1}$ and $\{y_2(n)\}_{n=mK}^{mK+N-1}$, respectively, where K is the overlap factor. The covariance matrix is composed of the frequency spectra, as follows:

$$Y(m,\omega_k) = [Y_1(m,\omega_k) \quad Y_2(m,\omega_k)]^T \tag{1}$$

$$R(m,\omega_k) = E[Y(m,\omega_k)Y^H(m,\omega_k)]. \tag{2}$$

We use the following formula to obtain the covariance matrix at time-frequency bin (m,ω_k).

$$\hat{R}(m,\omega_k) = \frac{1}{C}\sum_{l=m-C+1}^{m} Y(l,\omega_k)Y^H(l,\omega_k) \tag{3}$$

where C is the number of time-blocks averaged to obtain the short-time estimate. The magnitude-squared coherence

(MSC) function is calculated from this covariance matrix as follows:

$$T_{coh}^{(i,j)}(m,\omega_k) = \frac{|[R(m,\omega_k)]_{ij}|^2}{[R(m,\omega_k)]_{ii}[R(m,\omega_k)]_{jj}} \qquad (4)$$

where ii, jj and ij referes to the elements of the matrix R. Since this value approaches 1 in the interval for a single sound source, the interval of a single sound source can be detected.

2.2 Method for locating a single sound source

Here, the estimation approach to horizontal direction and the vertical direction that we proposed before [3] is described. However, the sound source is limited to a white noise in this technique. This time, the technique for enhancing it to the voice is described.

2.2.1 Localization of the horizontal plane

The results of a psychology experiment appear to indicate that humans presume the direction of a sound source in a horizontal plane using primarily ITD and ILD as the clue. When the sound source distance is 50 cm or more, ITD becomes approximately constant for the direction of the sound source[4]. Humans use the time difference and the acoustic pressure difference as the clue of determining the position of a sound source on a horizontal plane. In the present study, only the ITD information is used as an indicator to estimate the horizontal direction. The ITD is calculated from the sampling rate f_s and difference Δt of the specimen point at which the cross-correlation coefficient value of the right-hand and left-hand sound data reaches the maximum, becoming $\Delta t/f_s$. The direction of the sound source is estimated from this ITD. The distance difference l of arrival is calculated from the ITD and the sound velocity c by the equation $l = c\Delta t/f_s$. When the distance between the sound source and the microphones is great, the direction of the sound source is calculated using the difference distance of arrival as $\theta = \arccos(l/d)$. Where d is the distance between the microphones.

2.2.2 Localization of the vertical plane

Humans use single-ear cues induced by the pinna and other parts of the head in order to detect the vertical direction of a sound and determine whether the sound originated from the front or the rear. The pinna causes a change in the spectrum pattern depending on the direction of the sound source and hence can be used as the cue for direction estimation. The more the sound source contains the frequency element in the large range, the more the change of the spectrum takes place in the large range. Therefore, the amount of information on the direction of the sound source can be increased, and the accuracy of the estimated direction of the sound source is increased. In contrast, localization becomes difficult because little information is obtained when the sound source is a pure tone. According to the results of a psychological experiment conducted by Hebrank, humans use information on sound in a comparatively high-frequency area (4 - 15 kHz) as the localization cue [5]. In addition, in a measurement using an artificial ear, Shaw et al. showed that two or more dips, depending on the direction of the sound source, existed in its spectrum [6]. Ono et al. realized a single-ear sound source localization system with a reflector [7] that can localize the direction of a sound source of white noise. The reflector causes spectrum pattern changes depending on the direction of the sound source and is designed so that the logarithm of the frequencies of the local minima of the spectrum increase in proportion to the direction of the sound source.

Humans use a single ear, including the pinna, etc., to judge the vertical direction and forward-back (i.e., horizontal) direction of a sound. The pinna enables the human to estimate the direction of the sound source by detecting the change in the spectrum of the sound, which depends on the direction of the sound source.

We constructed a robotic system that learns sound localization in the vertical direction using local minimum frequencies [3]. In a related study, Ono et al. [7] estimated the vertical direction using only a small frequency band that included only one local minimum. In contrast, the proposed system uses a larger frequency band and acquires the localization capability by learning based on a self-organization feature map. More specifically, Ono et al. designed the reflector so that the logarithm of the local minimum frequency and the direction of the sound source had a linear relationship and the frequency band contained one local minimum, whereas the proposed system uses a wider frequency band that contains two or more local minimum frequencies and acquires the relationship between the local minimum frequency and the direction of the sound source by learning .

The proposed system uses a feature vector to handle two or more minimum values. Feature vector I is generated from the position of the minimum point frequency at each sound source position, as shown in Figure 1. The upper graph in this figure shows the frequency response, and the lower graph shows the feature vector. When there are N minimum values, the frequency at which a minimum value is taken is assumed to be $f_1, f_2, \dots f_N$, the range of the frequency used is defined as $f_{\min} \sim f_{\max}$. The i-th element I_i(i=1,...,M), for the case in which the dimension of the feature vector is assumed to be M, is defined as

$$\begin{aligned} I_i &= \sum_{j=1}^{N} \exp\left(-\frac{(i-p_j)^2}{\sigma^2}\right) && (5) \\ p_j &= \frac{f_j - f_{min}}{f_{max} - f_{min}} M. && (6) \end{aligned}$$

Here, we use $\sigma = 4$, $f_{min} = 5,000$, $f_{max} = 20,000$, and $M = 60$. The feature map is modeled as a Kohonen self-organizing feature map with the feature vector as input. The input data and the relation of the sound source position are learned by applying the lookup method to the feature map table. This system can acquire a robust ability to detect changes in the environment by comparing single source information. In the present study, the sound source localization in the vertical direction is estimated using the proposed technique.

2.2.3 Correspondence to any sound source

Since the technique that uses such a reflector for vertical localization requires white noise to be used as the sound source, this technique cannot correctly locate a source in the vertical direction when sound sources other than white noise, e.g., voices, are used. In order to obtain the transfer function of the reflector, except when the sound source has a flat magnitude property, that is, white noise, the magnitude property of the sound source is needed.

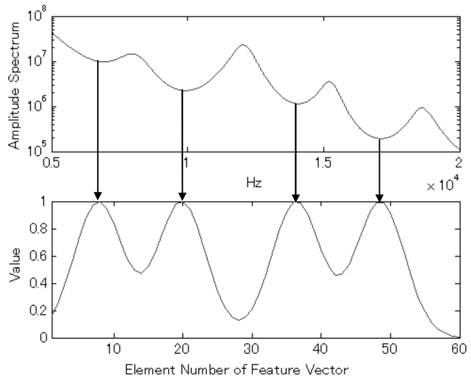


Figure 1: Generation of the feature vector

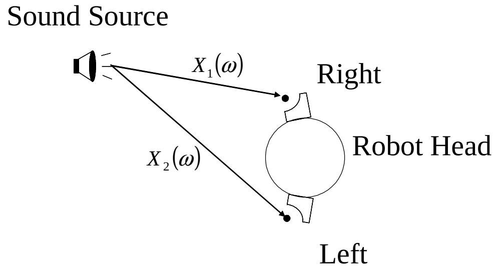


Figure 2: Influence of the reflector

In order to solve this problem, we propose a technique of using microphone information other than that used for localization. Figure 2 shows the arrangement of the microphones and the reflector. The reflector does not influence the sound ($X_2(\omega)$) that enters a microphone from a distant sound source, but, because of the shape of the reflector, does influence the sound ($X_1(\omega)$) that enters the microphone from a nearby sound source. Here, the magnitude properties of the sound source is denoted by $S(\omega)$, the transfer functions from a sound source to the right-hand and left-hand microphones are denoted by $H_1(\omega)$ and $H_2(\omega)$, respectively, the transfer function of the reflector is expressed as $R(\omega)$, and the magnitude properties of the observation signal of the microphones are denoted by $X_1(\omega)$ and $X_2(\omega)$, which are obtained as follows:

$$X_1(\omega) = R(\omega)H_1(\omega)S(\omega) \quad (7)$$

$$X_2(\omega) = H_2(\omega)S(\omega). \quad (8)$$

From this expression, the magnitude property of the reflector can be obtained as follows:

$$|R(\omega)| = \frac{|H_2(\omega)||X_1(\omega)|}{|H_1(\omega)||X_2(\omega)|.} \quad (9)$$

We assume that the difference in the magnitude property between the left and right ears can be neglected. That is, when the relationship for the magnitude properties $|H_1(\omega)|$ and $|H_2(\omega)|$ between the sound source and the microphone is assumed to be $H_1(\omega)| = |H_2(\omega)|$, the magnitude property of

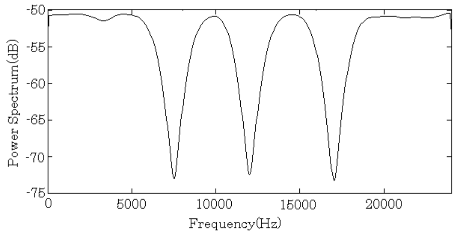
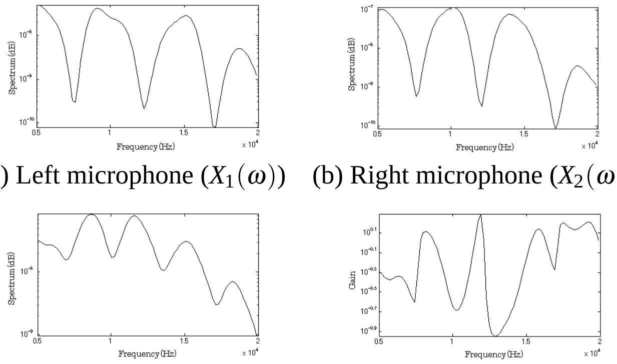


Figure 3: Frequency characteristic of the test sound



(a) Left microphone ($X_1(\omega)$) (b) Right microphone ($X_2(\omega)$)

(c) Amplitude (d) Results ($R(\omega)$)

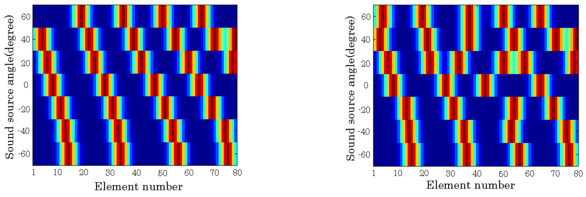
Figure 4: Frequency characteristics

the reflector becomes

$$|R(\omega)| = \frac{|X_1(\omega)|}{|X_2(\omega)|.} \quad (10)$$

Thus, it becomes possible to calculate the magnitude property $|R(\omega)|$ of the reflector using only the observation signal for the right-hand and left-hand microphones.

We conducted an experimented to verify the validity of the proposed technique (10). We used a sound source having three valleys on the frequency axis, as shown in Fig. 3. Figure 4(a) shows the magnitude property ($X_1(\omega)$) for the microphone on the side near the sound source, and Fig. 4(b) shows the magnitude property ($X_2(\omega)$) for the microphone on the side far from the sound source. In addition, Fig. 4(c) shows the magnitude property between the sound source and the microphone for the case in which white noise is used as the sound source. In an ideal environment that does not include attenuation by an echo or propagation, this magnitude property becomes the magnitude response of the reflector. Finally, Fig. 4(d) shows the calculated magnitude property ($|R(\omega)|$). Although the positions of the valleys in Fig. 4(c) and Fig. 4(d) are close to each other, the shapes are considerably different . The feature map was generated while changing the height of the test sound source for the purpose of comparison with the feature map [3] generated by white noise. Figure 5 shows the results. Figure 5(a) shows the feature map generated by white noise, and Figure 5(b) shows the feature map generated by the sound shown in Fig. 3, which is similar to the white noise. The proposed technique is thought to enable localization similar to that used for white noise. In summary, the vertical direction of the voice can be estimated using the proposed technique.



(a) White noise (b) Test sound

Figure 5: Feature maps generated by feature vectors

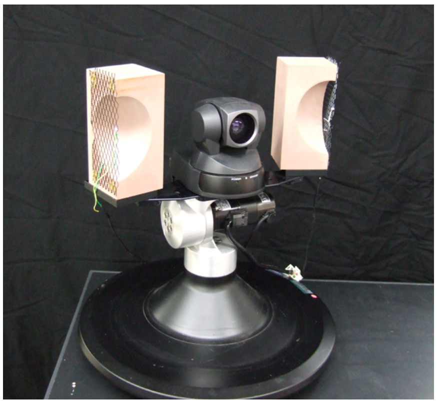


Figure 6: Robot head incorporating two microphones with reflectors as pinnae

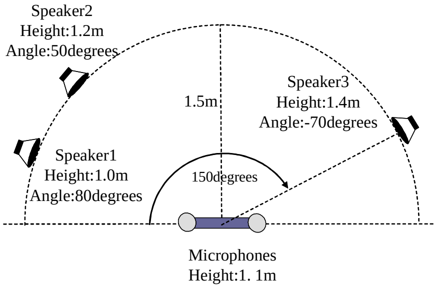


Figure 7: Arrangement of microphones and speakers

3. EXPERIMENT ON THE LOCALIZATION OF TWO OR MORE SOUNDS

3.1 Experimental environment

To demonstrate the effectiveness of the proposed method, we conducted the experiment in an anechoic room. Two omnidirectional microphones with reflectors were attached to the robot (Fig. 6). The height from the ground to the microphone was 1.1 m, and the distance between microphones was 30 cm. Three sound sources were placed at three heights spaced at intervals of 20 cm, and the proposed system can distinguish the three heights. The three speakers were arranged as shown in Fig. 7.

3.2 Experimental results

Figure 8 shows the three source signals used in the experiment. The shadow in this figure is the one main sound source. The values of the parameters used when the value of the MSC function was obtained are N = 1,024 and K = 820. Moreover,

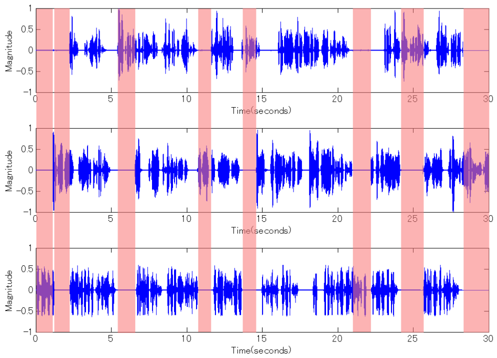


Figure 8: Three source signals used in the experiment

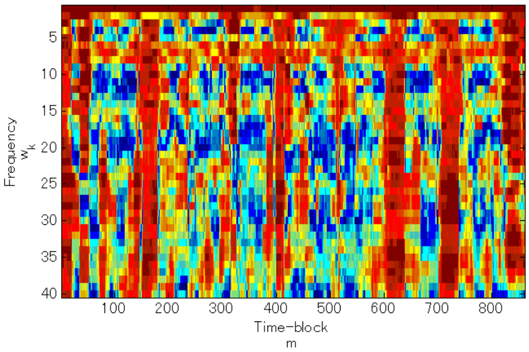


Figure 9: MSC value

the number of time-blocks, C, is 15. Figure 9 shows the value of the MSC function of the two audio signals obtained from the microphones. Figure 10 shows the sum of the MSC functions. This figure represents the sum of the function values of 31 points of the discrete frequency from k = 10 to k = 40. The function value is large in one part of one sound source. Figure 11 shows the results for localization in the horizontal direction for the one detected sound source interval. This figure shows that an approximately correct direction of the sound source can be estimated. Moreover, the three heights in the vertical direction are correctly distinguished. The feature vectors of Equation 5 are show in Figure 12. The feature maps are shown in Figure 13. The top figure shows the feature vector and feature map to speaker 1. The center figure shows the feature vector and feature map to speaker 2. The bottom figure shows the feature vector and feature map to speaker 3. The feature vector and the feature map have good correspondence, although a failure in the detection of a minimum value occurred.

3.3 Discussion

When an interval of a single sound source can be detected accurately from among multiple sound sources by MSC, the proposed method is similar to the sound localization technique used for a single sound source. Horizontal accuracy is one sample or less. When this value is converted into the angle value, it is about two degrees in the vicinity of the front, ten degrees in the vicinity of both ends.

The proposed sound localization techniques [8][9] for

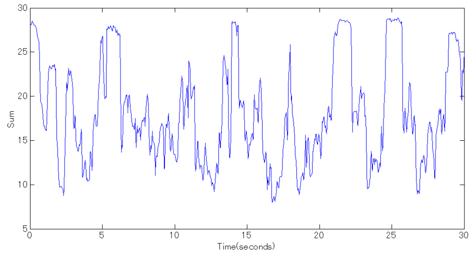


Figure 10: MSC

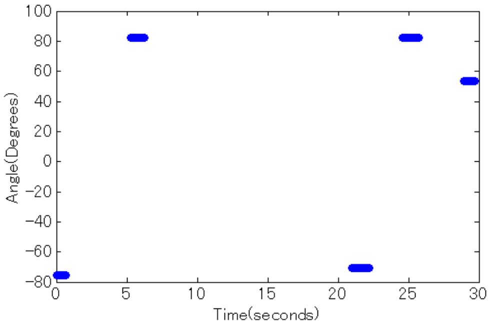


Figure 11: Results of localizing in the horizontal direction

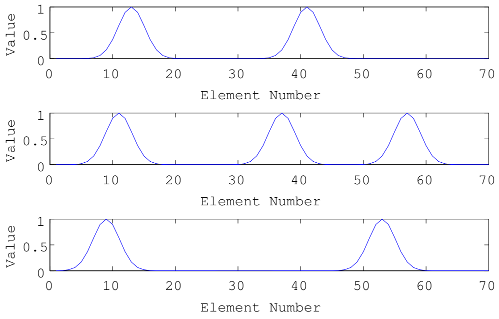


Figure 12: Feature vector to each sound source direction

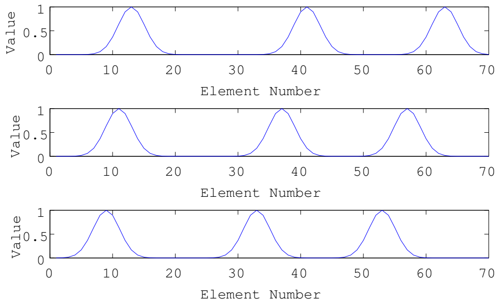


Figure 13: Feature map to each sound source direction

multiple sound sources using two or more microphones were thus validated. For the technique using the main-lobe fitting method, the average horizontal location error margin for three sound sources was five degrees or less, which is more accurate than the proposed method.

4. CONCLUSION

In the present study, we have proposed a technique for estimating the direction of the sound source of a voice from three different positions and have demonstrated the effectiveness of the proposed technique experimentally. It was possible to estimate the position of the horizontal direction and the vertical direction of three sound sources using only two microphones by the proposed method.

In the proposed method, the feature of the sound source is denied by assuming the acquisition sound on the side that reaches the microphone directly to be the sound source. As a result, the localization of the vertical direction became possible for an arbitrary sound. However, the proposed technique does not work correctly for a sound source in the vicinity of the front for the arrangement of the reflector used in the present study. The reason for this is that the sound originating from the vicinity of the front does not pass the reflector right and left both, or to pass the reflector right and left both. The solution to this problem will be examined in the future.

REFERENCES

[1] N. Ono, S. Ando, Sound Source Localization Sensor with Mimicking Barn Owls, in: Transducers'01, 2001, pp. 1654–1657.

[2] S. Mohan, M. L. Kramer, B. C. Wheeler, D. L. Jones, Localization of nonstationary sources using a coherence test, in: 2003 IEEE Workshop on Statistical Signal Processing, 2003, pp. 470–473.

[3] H. Nakashima, T. Mukai, 3D Sound Source Localization System Based on Learning of Binaural Hearing, in: Proc. IEEE SMC 2005, 2005, pp. 3534–3539.

[4] D. S. Brungart, W. M. Rabinowitz, Auditory localization of nearby sources. Head-related transfer functions, J. Acoust. Soc. Am. 106 (3) (1999) 1465–1479.

[5] J. Hebrank, D. Wright, Specral cues used in the localization of sound sources on the median plane, J. Acoust. Soc. Am. 56 (6) (1974) 1829–1834.

[6] E. A. G. Shaw, R. Teranishi, Sound Pressure Generated in an External-Ear Replica and Real Human Ears by a Nearby Point Source, J. Acoust. Soc. Am. 44 (1) (1968) 240–249.

[7] N. Ono, Y. Zaitsu, T. Nomiyama, A. Kimachi, S. Ando, Biomimicry Sound Source Localization with Fishbone, T.IEE 121-E (6) (2001) 313–319.

[8] Y. Sasaki, et al, 2D Sound Localization from Microphone Array Using a Directional Pattern, in: The 25th Annual Conference of The Robotics Society of Japan,, 2007.

[9] K.Nakadai, H.Okuno, H.Kitano, Speech localization, separation and recognition by active audition for humanoid, in: Proc. of the 16th Meeting of Special Interest Group on AI Challenges, 2002, pp. 25–32.