Reliable Speaker Localization using Signal-to-Noise Ratio Information in Noise Environments for the SIG-2 Humanoid Robot

*Ui-Hyun KIM and Hiroshi G. OKUNO (Kyoto University)

1. Introduction

Speaker localization is one of the most important techniques to achieve more natural and intelligent human-robot interaction (HRI). This is because robots are able to watch the directions of talkers to express their interest in the conversation by their speaker localization systems. “Binaural” literally means having or relating to two ears. For robots, this means only two microphones located in the left and right sides of the robot head like human ears. Recently, research on binaural robot audition has been growing in part because the cost of binaural audition hardware using only two microphones is much cheaper and uses much less computational power than multi-channel audition devices. Moreover, research on binaural audition can contribute on understanding the human hearing mechanism [1]. For these reasons, the binaural speaker localization system is becoming necessary for humanoid robots.

The primary clues for sound localization have been discovered by a number of researchers, including the inter-aural level difference (ILD), the inter-aural time difference (ITD), and the spectral modification due to the pinna, head, shoulder and torso. These clues are contained in the head related transfer function (HRTF) [2]. The ITD, more commonly referred to as the time difference of arrival (TDOA), plays an important role in sound localization; the sound signals arrive at each microphone at different times due to the finite speed of sound and different positions of microphones from the sound source. One of the most popular algorithms using this ITD clue is the generalized cross-correlation method (GCC) with its phase transform (PHAT) weighting [3].

Many robot audition systems have been developed using the GCC-PHAT method, and their performance has gradually improved. However, their common problem is that the localization performance is degraded when target sound sources are corrupted by additive noise. In practical localization situations, background noise is generated by several specific noise sources, such as motor of the robot, air conditioning and computer fans. These noise sources are spectrally correlated with the target sound sources and these correlations affects the localization performance.

To cope with this problem, we propose a spectral weighting function based on the signal-to-noise ratio (SNR) of each frequency bin for the GCC-PHAT method. This spectral weighting function helps the GCC-PHAT method to calculate with clean frequency bins of target sound sources by giving a low weight to the frequency bins corrupted by noise. This spectral weighting function was implemented and evaluated experimentally with our binaural DOA estimation method for the SIG-2 humanoid robot [4].

The paper is organized as follows. Section 2 summarizes the DOA estimation using the GCC-PHAT method for binaural robot audition. Section 3 presents a spectral weighting function for the GCC-PHAT method. Section 4 evaluates experimental results. Section 5 concludes this paper.

2. Binaural DOA estimation

In this section, we summarize the ML-based DOA estimation using the GCC-PHAT method with a time delay factor to compensate for multipath interference due to the diffraction of sound wave along the shape of the robot head.

2.1 ML-based DOA estimation

This paper uses a time-frequency domain approach with a F-point short-time Fourier transform (STFT) and a far-field assumption. Since sound signals consist of varied changes in volume, timbre, or tone over time, and since they usually occur far from the position of microphones in the practical localization situation, the STFT and the far-field assumption have been generally used for the DOA estimation [4].

The observed signals from the left and right microphones in a situation with K-sound sources can be mathematically modeled as

$$\begin{aligned} X_l[f,n] &= \sum_{k=1}^{K} \alpha_{lk}[f] S_k[f,n] e^{-j2\pi \frac{f}{F} fs \tau_{lk}} + N_l[f,n] \\ X_r[f,n] &= \sum_{k=1}^{K} \alpha_{rk}[f] S_k[f,n] e^{-j2\pi \frac{f}{F} fs \tau_{rk}} + N_r[f,n], \end{aligned} \tag{1}$$

where $X_{l,r}[f,n]$, $S_k[f,n]$, and $N_{l,r}[f,n]$ are the f-th elements of the STFT of the measured signals from the two microphones l and r, the k-th sound sources, and additive noise, respectively, on the n-th time-frame index. $f \in \{1, \ldots, F\}$ denotes a frequency bin, F is the time-frame size of the STFT, and fs is the sampling frequency. $\alpha_{lk,rk}$ and $\tau_{lk,rk}$ are the attenuation factor and time delay from the position of the k-th sound source to each microphone, respectively.

The Maximum Likelihood (ML)-based DOA estimation for multiple sound sources is basically defined by the GCC-PHAT method in the frequency domain with a threshold η_{DOA} ranging from 0 to 1, as follows:

$$\begin{aligned}&\text{if} \quad \hat{P}_{\theta}[n] > \eta_{DOA} \quad \text{and} \quad \theta \text{ has a peak} \\ &\text{then} \quad \theta \in \hat{\theta}_{mle_k} \qquad ,\end{aligned} \tag{2}$$

where

$$\hat{P}_{\theta}[n] = \frac{1}{F}\sum_{f=1}^{F} G^{PHAT} X_l[f,n] X_r^*[f,n]\; e^{j2\pi\frac{f}{F}fs\tau_{lr}(\theta)}, \tag{3}$$

$$G^{PHAT} = \frac{1}{\left|X_l[f,n]X_r^*[f,n]\right|}, \tag{4}$$

$$\tau_{lr}(\theta) = \frac{d_{lr}}{v}\sin\left(\frac{\theta}{180}\pi\right), \tag{5}$$

θ∈{-90°, ..., +90°} is an angle of sound incidence, * is the complex conjugate, G^{PHAT} is a normalization factor to preserve only the phase information, τ_{lr} is the inter-aural TDOA defined by the relationship: $\tau_{lr} = \tau_{rk} - \tau_{lk}$ assuming microphone *l* as a reference, d_{lr} is the distance between two microphones, and *v* is the speed of sound (340.5 m/s, at 15 °C, in air). Since the estimated DOAs of multiple sound sources in (2)-(5) are obtained by finding each expected angle of sound incidence *θ* that makes each peak with a threshold η_{DOA} ranging from 0 to 1 in the frequency domain instead of the conventional estimation using the cross-power spectrum phase (CSP) analysis in the time domain [6], this ML-based DOA estimation has two advantages contributing to improved accuracy:

1. Selective resolution of the DOA estimation by adjusting the interval of the angles of sound incidence.
2. Individual calculations on frequency bands by summing only the frequency bins of interest in the cross-power spectrum, e.g. summing the frequency bins from around 60 Hz to 7000 Hz for the human voice.

2.2 New TDOA factor for binaural robot audition

Sound wave easily bends and spreads along the shape of the robot head, and these attributes of sound cause different TDOAs from a sound source along the front-head path and the back-head path with multipath interference in binaural robot audition. We derived a new TDOA formula for the TDOA factor τ_{lr} in (5) that takes into account this diffraction of the sound wave around the robot head with multipath interference, which is defined as:

$$\begin{aligned}\tau_{lr}(\theta) =& \frac{d_{lr}}{2v}\left\{\frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right\} \\ &- \frac{d_{lr}}{2v}\left\{\mathrm{sgn}(\theta)\pi - \frac{2\theta}{180}\pi\right\}\left|\beta_1 \sin\left(\frac{\theta}{180}\pi\right)\right|.\end{aligned} \tag{6}$$

where β_1 is an attenuation factor and *sgn* is the signum function that extracts the sign of *θ*, i.e. if *θ* has the negative sign, then *sgn(θ)* will be -1.

3. SNR-weighted DOA estimation

In the practical localization situation, the noise sources are spectrally correlated with the target sound sources and they usually distribute in specific narrow frequency bands. This means that SNR between target signals and additive noise will be large in frequency bands of clean target signals. Based on this characteristic of noise, we propose a spectral weighting function based on SNR information, which gives different weights to each frequency bin. To derive robust SNR, we employee the SNR estimation technique used in the two-step noise reduction (TSNR) method [7] with recursive noise adaption.

3.1 Spectral weighting function

In the first step in the TSNR technique, the *a priori* SNR is computed with the decision-directed (DD) estimation approach to reduce the bias of an estimator [8] as follows:

$$\begin{aligned}\hat{\xi}_{DD}[f,n] = \beta_2 &\frac{\left|\hat{S}[f,n-1]\right|^2}{\lambda_N[f,n]} \\ &+ (1-\beta_2)P\{\gamma[f,n]-1\},\end{aligned} \tag{7}$$

where β_2 is the forgetting factor whose value is typically chosen as 0.98 (0< β_2<1), *S^[f,n]* and $\lambda_N[f,n]$ are the estimated sound sources and the noise variance, respectively, *P[·]* is the half-wave rectification which is defined by *P[x]=x* if *x≥0* and *P[x]=0* otherwise, and $\gamma[f,n]=|X[f,n]|^2/\lambda_N[f,n]$ is the *a posteriori* SNR. Then, the spectral gain $G_{DD}[f,n]$ is obtained by applying Equation (7) to the Wiener amplitude estimator as follows:

$$G_{DD}[f,n] = \frac{\hat{\xi}_{DD}[f,n]}{1+\hat{\xi}_{DD}[f,n]}. \tag{8}$$

In the second step, $G_{DD}[f,n]$ is used for estimation of the TSNR *a priori* SNR as follows:

$$\hat{\xi}_{TSNR}[f,n] = \frac{\left|G_{DD}[f,n]X[f,n]\right|^2}{\lambda_N[f,n]}. \tag{9}$$

Finally, the proposed spectral gain $G_{SW}[f,n]$ to weaken the influence of noise is obtained by applying Equation (9) to the Wiener amplitude estimator again:

$$G_{SW}[f,n] = \frac{\hat{\xi}_{TSNR}[f,n]}{1+\hat{\xi}_{TSNR}[f,n]}, \tag{10}$$

and the estimated sound sources to be used in (7) can be obtained by applying $G_{SW}[f,n]$ to the input signal as the following equation:

$$\hat{S}[f,n] = G_{SW}[f,n]X[f,n]. \tag{11}$$

3.2 Noise adaptation

To compensate for fluctuations of noise power level, the noise variance $\lambda_N[f,n]$ is updated in a recursive way as follows:

$$\lambda_N[f,n] = \beta_3\lambda_N[f,n-1] \\ +(1-\beta_3)\left\{\left|X[f,n]\right|^2 - |\hat{S}[f,n]|^2\right\}, \tag{12}$$

where β_3 is the forgetting factor whose value is typically chosen as 0.98 ($0 < \beta_3 < 1$).

3.3 DOA estimation with spectral weighting function

The reliable DOA estimation in noisy environments is can be derived from (3), (4), and (6) with the spectral weighing function (10) and the noise adaptation process (12) as follows:

$$\hat{P}_\theta[n] = \frac{1}{F}\sum_{f=1}^{F} G_{SW}\frac{X_l[f,n]X_r^*[f,n]}{\left|X_l[f,n]X_r^*[f,n]\right|}e^{j2\pi\frac{f}{F}fs\tau_{lr}(\theta)}. \tag{13}$$

4. Experimental results

We evaluated the proposed spectral weighting function in single- and three-speaker situations to verify that it can make fewer localization errors in noisy environment. For this purpose, the method was implemented in the binaural audition system of our SIG-2 humanoid robot as shown in Fig. 1.

The experimental room with reverberation time of about 120 milliseconds contained background noise from air conditioners, personal computers, and music. The average sound pressure level (SPL) of background noise was about 32.7 dB and those of target speech signals were about 56.2 dB. The SIG-2 humanoid robot was placed in the center of the room and tested at a 1.5 ~ 2.5 m radius.

The values for the threshold η_{DOA} in (2) and the attenuation factor β_1 in (6) of the ML-based DOA estimation were set at 0.1 and 0.2, and the forgetting factors β_2 in (7) and β_3 in (12) were set at 0.98, respectively. In addition, the system recorded background noise for 2 seconds before operating the SIG-2 humanoid robot to estimate the initial noise variance in (7), (9), and (12). Then the noise adaptation function is performed on the speech-absent frames determined by the VAD decision rule [9] in our binaural robot audition system.

Figure 2 shows the root mean square error (RMSE) of the experimental results with and without the spectral weighting function on a single-speech occasions at each locus of the azimuth from -90 degrees to +90 degrees in 10-degrees-units. According to Fig. 2, using the spectral weighting function could reduce the average RMSE of the DOA estimations by about 1.5 degrees.

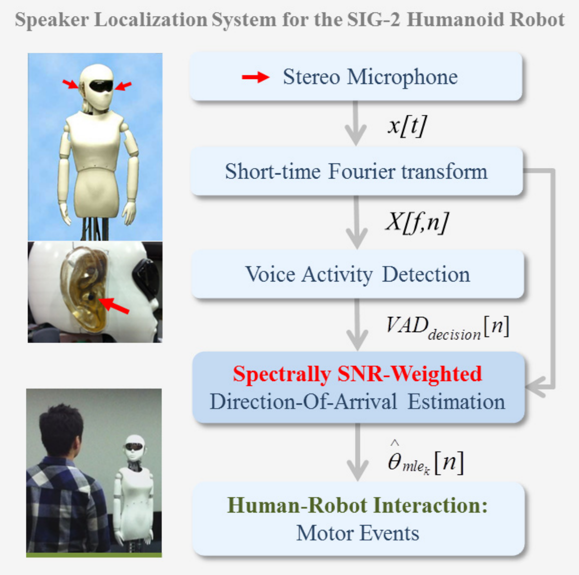
Speaker Localization System for the SIG-2 Humanoid Robot
Stereo Microphone
x[t]
Short-time Fourier transform
X[f,n]
Voice Activity Detection
$VAD_{decision}[n]$
Spectrally SNR-Weighted
Direction-Of-Arrival Estimation
$\overset{\wedge}{\theta}_{mle_k}[n]$
Human-Robot Interaction:
Motor Events

Fig.1 Flowchart of the speaker localization system for the SIG-2 humanoid robot.

Figures 3 show the experimental results of three-speaker localization for 6 seconds. As shown in Fig. 2-(c) and Fig. 2-(d), the spectral weighting function could help the ML-based DOA estimation to produce a more reliable and accurate multiple DOA estimations.

5. Conclusion

In this paper, we presented a spectral weighting function based on SNR information for the GCC-PHAT method to cope with the problem that the localization performance is degraded when target sound sources are corrupted by additive noise. Based on the characteristic that SNR between target signals and additive noise is large in frequency bands of clean target signals, we derived the spectral weighting function by using SNR information to gives different weights to frequency bins in calculating of the GCC-PHAT method.

Experimental results demonstrated that the DOA estimation with the proposed spectral weighting function can produce a more reliable and accurate multiple estimations in noisy environments.

For the future work of this study, we are planning to improve our spectral weighting function to handle the ambiguity of speaker identification in multiple DOA estimations by applying independent vector analysis (IVA) [10].

6. Acknowledgment

This research was partially supported by JSPS Grant-in-Aid for Scientific Research (S), JST Japan-France Cooperative Research Project BINAAHR, and the Global COE Program of Graduate School of

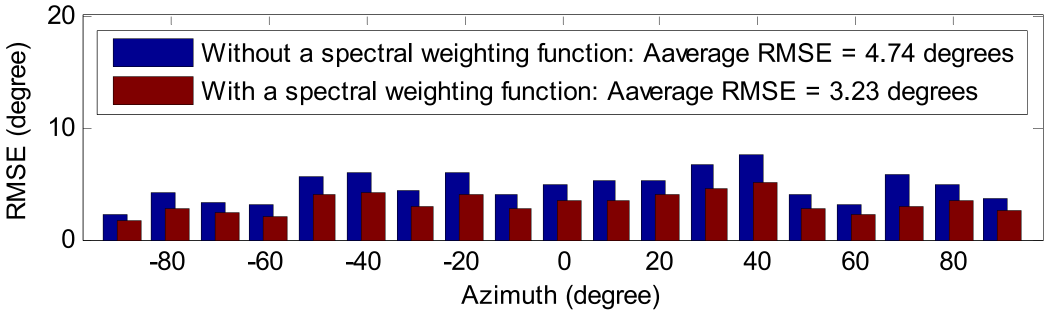

Fig.2 RMSE of experimental results in speaker localization.

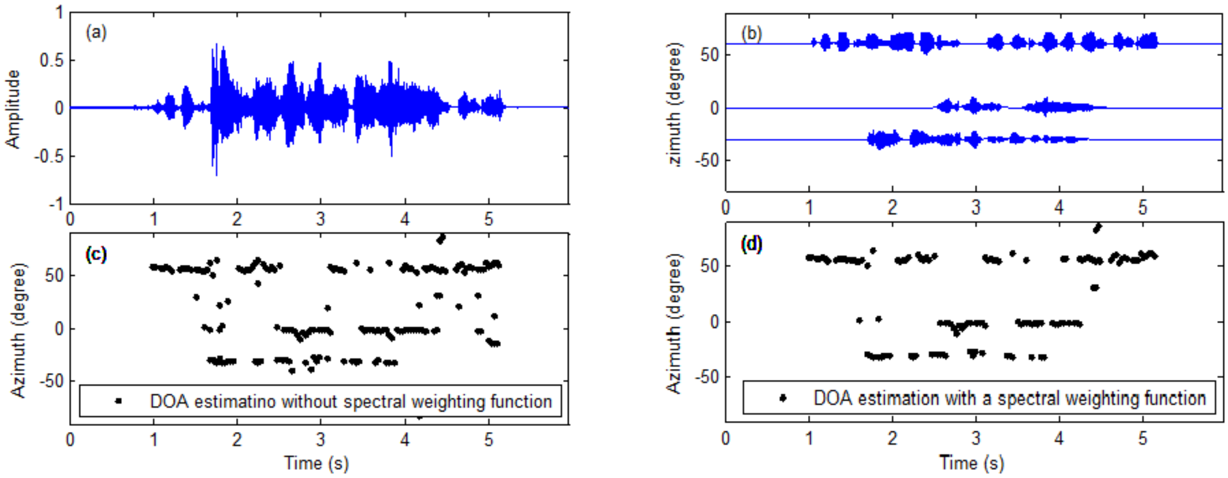

Fig.3 Results of three-speaker localization (a) Input signal consisting of two male speeches and a female speech on the left microphone. (b) Real directions and spoke times of three speakers. (c) Results of DOA estimation without the proposed spectral weighting function. (d) Results of DOA estimation with the proposed spectral weighting function.

Informatics, Kyoto University.

REFERENCES

[1] J. Blauert and J. Braasch, "Binaural Signal Processing," in Proc. IEEE Int. Conf. on Digital Signal Processing (DSP), pp. 1-11, Greece, July 2011.

[2] C. I. Cheng and G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," J. Audio Eng. Soc., vol. 49, pp. 231–249, Apr. 2001.

[3] 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, no. 4, pp. 320-327, 1976.

[4] U. H. Kim and H. G. Okuno, "Robust Localization and Tracking of Multiple Speakers for Binaural Robot Audition," in Proceeding of the 2012 IEEE-RAS International Conference on Humanoid Robots, Osaka, Japan, December 2012.

[5] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition). Cambridge, MA: MIT Press, 1997.

[6] M. Matassoni and P. Svaizer, "Efficient time delay estimation based on cross-power spectrum phase," in Proceeding of European Signal Processing Conference (EUSIPCO), Florence, Italy, September. 2006.

[7] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", IEEE Transactions on Audio, Speech & Language Processing, pp.2098-2108, 2006.

[8] Y. Ephraïm and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Transactions on Acoustics, Speech, Signal Processing, vol. no. 6, pp. 1109–1121, Dec. 1984.

[9] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 365–368, 1998.

[10] T. Kim, T. Attias, and S. Y. Lee, "Blind Source Separation Exploiting Higher-Order Frequency Dependencies," IEEE Transactions on Audio, Speech & Language Processing, vol. 15, no. 1, pp. 70-79, Jan. 2007.