4C-1

Speaker Localization Using Two-Channel Microphone on the SIG-2 Humanoid Robot

Ui-Hyun Kim† Toru Takahashi† Tetsuya Ogata† Hiroshi G. Okuno†

Graduate School of Informatics, Kyoto University†

1. INTRODUCTION

Speaker localization is one of the most important techniques to achieve natural and intelligent human-robot interaction (HRI) because robots need to 1) identify the direction of a talker through the measurements of the acoustic signals from microphones, and 2) watch at the position of a talker for notifying that they are now ready to receive an order or express their interest in conversation. Moreover, speaker localization with two-channel microphones is required for humanoid robots by two reasons; 1) cost reduction and 2) portability. The cost of stereo input devices is much cheaper than that of multi-channel analog-to-digital (AD) devices. Moreover, the market of binaural audition hardware is economically growing. It can be easily embedded on PCs, TVs and other ICT devices. Thus, the development of SL with two microphones is a necessary for robotics [1].

Speaker localization is usually based on voice activity detection (VAD) and sound source localization (SSL). The one of the most popular algorithms for SSL is the generalized cross-correlation method (GCC) and its phase transform (PHAT) weighting. It is well known that GCC-PHAT method performs very well in noise and reverberation environments [2]. Many robot audition systems have been developed with GCC-PHAT method and their performance has generally improved more and more. However, the following two issues of the conventional GCC-PHAT method should be considered and improved: 1) Diffraction of sound wave by the robot's head in non-free space. 2) Low performance around the lateral direction of sound source with one pair of microphones. In this paper, we describe these two issues as a problem and address their solutions: 1) The formula considered the diffraction of sound wave after assuming that the shape of the robot's head is a circle. 2) Applying the maximum likelihood (ML)-based direction-of-arrival (DOA) estimator in frequency domain. These solutions implemented and evaluated with experimental results in our speaker localization system using two-channel microphone embedded on the SIG-2 Humanoid robot.

2. DIRECTION-OF-ARRIVAL ESTIMATION

This paper employs a time-frequency domain approach with a *T*-point short-time Fourier transform (STFT). The received signal from *m*-th microphone can be mathematically modeled as

$$X_m(f,n) = \alpha S(f,n) + N_m(f,n) \quad (1)$$

where $X_m(f,n)$, $S(f,n)$, and $N_m(f,n)$ are *f*-th elements of STFTs of measured signal from *m*-th microphone, sound source, and uncorrelated additive noise respectively, on *n*-th time-frame index. α is an attenuation factor, $f \in \{0, fs/T, \cdots, fs(T-1)/T\}$ is a frequency, *fs* is the sampling frequency, and *T* is the frame size for STFT.

2.1 CONVENTIONAL GCC-PHAT METHOD

GCC-PHAT method to estimate the time difference of arrival (TDOA) τ_{ij} between two microphones *i* and *j* is derived [3] by

$$R_{x_i x_j}(f,n) = \sum_{f=0}^{f_s(T-1)/T} G^{PHAT}(f,n) X_i(f,n) X_j^*(f,n)\, e^{j2\pi f\tau}, \quad (2)$$

$$csp_{ij}(t,n) = ISTFT[R_{x_i x_j}(f,n)], \quad (3)$$

$$\hat{\tau}_{ij}(n) = \arg\max_t (csp_{ij}(t,n)) \quad (4)$$

where

$$G^{PHAT}(f,n) = \frac{1}{\left| X_j(f,n) X_i^*(f,n) \right|}, \quad (5)$$

Rx_ix_j is the cross-correlation function, * is the complex conjugate, csp_{ij} is the coefficient of the cross-power spectrum phase analysis (CSP), *t* is the time index, and *ISTFT* is the inverse short-time Fourier transform.

After TDOA τ_{ij} is estimated, the sound source direction is derived from the following equation.

$$\hat{\theta}(n) = \sin^{-1}\left(\frac{\hat{\tau}_{ij}(n)c}{d_{ij} fs}\right)\frac{180}{\pi} \quad (6)$$

where $\hat{\theta}$ is an estimated direction of a sound source, d_{ij} is the distance between two microphones, and *c* is the sound speed (340.5 m/s, at 15 °C, in air).

2.2 PROBLEM

This conventional DOA estimation using GCC-PHAT method has two problems:

1) It has not considered the diffraction of sound wave when microphones are not located in a free space such as those in the head of robots or the inside of the ear.

2) It is restricted by the sampling frequency. Since the maximum value of Equation (3) exists in the time domain through ISTFT, τ_{ij} must depends on the sampling frequency of the signal. For example, if a sampling

† Ui-Hyun Kim, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno are with Speech Media Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan (e-mail: {euihyun, tall, ogata, okuno}@kuis.kyoto-u.ac.jp).

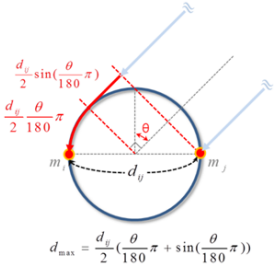

Figure 1. Simplified formula

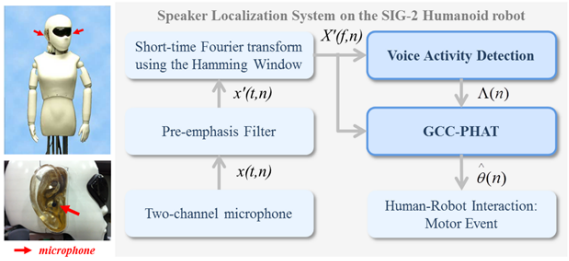

Figure 2. System flow chart for speaker localization

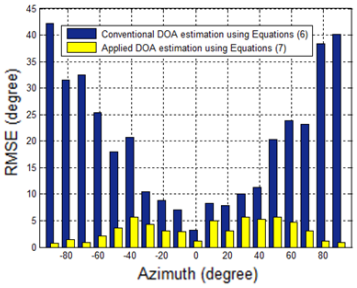

Figure 3. RMSE of Experimental results

frequency is 16 kHz, the minimum τ_{ij} that we can estimate is limited to 62.5 μsec (1 sec / 16 kHz). In other words, since the difference between TDOAs coming from -90 degrees and -80 degrees is less than 62.5 μsec with 15cm of d_{ij}, we cannot distinguish -90 degrees and -80 degrees.

These two problems have a great effect on the DOA estimation, especially around the lateral direction of a sound source coming from around ±90 degrees whose differences are less than the minimum τ_{ij}.

3. Improvement Of Sound Localization

The solutions to the two problems above are as in the following:

1) Consideration of the diffraction of sound wave. We applied the maximum delayed distance d_{max} between two microphones as d_{ij} factor in Equation (6) after assuming that the shape of the robot's head is a circle. Figure 1 shows the formula considered the diffraction of sound wave with the assumption of the circle shaped head of a robot and parallel sound incidence.

2) The ML-based DOA estimator in frequency domain. To solve the problem of unreliable DOA estimation at the left and right sides of the robot head, we applied the ML-based DOA estimator in frequency domain which can be easily derived from Equations (2)-(6) with d_{max} as follows:

$$\hat{\theta}(n) = \arg\max_{\theta} \sum_{f=0}^{f_s(T-1)/T} \frac{X_j(f,n)X_i^*(f,n)}{\left|X_j(f,n)X_i^*(f,n)\right|} e^{j2\pi\frac{f}{T}f_s\frac{d_{\max}}{c}} \quad (7)$$

where

$$d_{\max} = \frac{d_{ij}}{2}\left(\frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right). \quad (8)$$

This ML-based DOA estimator can be obtained by finding the maximum value from scanning expected degrees $\theta \in \{-\pi/2, \cdots, \pi/2\}$. Since this DOA estimator is calculated in frequency domain with the phase transform, it is able to slip the limitation of sampling frequency in time domain and have advantage of 1 degree resolution.

4. System and Experiments

Figure 2 shows the flow of our speaker localization system. As the body of our system, we employ VAD algorithm and GCC-PHAT method. Traditional VAD algorithms are usually designed using heuristics, which makes it difficult to optimize the relevant parameters. Thus, we employees a statistical model-based VAD algorithm proposed by Sohn et. al [4], where $\Lambda(n)$ is the likelihood ratio with a threshold as to whether the signal is voiced or unvoiced.

We tested the applied ML-based DOA estimation using our speaker localization system of the SIG-2 humanoid robot with air conditioner and personal computer noise and music. The noise dB was about 61.2. The male speaker was placed at each locus of 10-degrees-unit azimuth from -90 degrees to 90 degrees and 1.5m distance. The speech source occurred 10 times at each locus for the conventional DOA estimation and the applied DOA estimation. Figure 3 shows the root mean square error (RMSE) of the experimental results on 190 occasions (azimuth change: 19 times × speech: 10 times). We could verify that the applied ML-based DOA estimation in frequency domain reduces localization errors in comparison to those of the conventional DOA estimation.

5. Conclusion and Future Works

In this paper, we described two issues of the conventional DOA estimation based on GCC-PHAT method and their solutions. To reinforce system faculty of sound localization, we considered the diffraction of sound wave by the robot's head and applied the ML-based DOA estimator in frequency domain. Experimental results demonstrated that our applied method is effective and robust in spite of only using two microphones.

However, we could only localize the azimuth among ±90 degrees on the SIG-2 humanoid robot. In near future, we will first solve the problem of the front-back confusion using two microphones with analyzing the artificial pinna and HRTF for the entire azimuth localization.

References

[1] H. D. Kim et. al, "Binaural Active Audition for Humanoid Robots to Localize Speech over Entire Azimuth Range, Applied Bionics and Biomechanics, Special Issue on "Humanoid Robots"", Vol.6, Issue 3 & 4(Sep. 2009) pp.355-368, Taylor & Francis 2009.

[2] V. M. Trifa et. al, "Real-time Acoustic Source Localization in Noisy Environments for Human-robot Multimodal Interaction," IEEE Intl. on Robot and Human Interactive Communication, pp. 393-398, August. 2007.

[3] C. H. Knapp et. al, "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] J. Sohn et. al, "A Statistical Model-Based Voice Activity Detection," IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1-3, 1999.