Sound Localization Based on Excitation Source Information for Intelligent Home Service Robots

Keun-Chang Kwak

Dept. of Control, Instrumentation, and Robotic Engineering, Chosun University
375 Seosuk-dong, Dong-gu, Gwangju, 501-759, Korea
kwak@chosun.ac.kr

Abstract. This paper is concerned with Sound Localization (SL) using Excitation Source Information (ESI) and effective angle estimation for intelligent home service robots that are equipped with multi-channel sound board and three low-cost condenser microphones. The main goal is to localize a caller by estimating time-delay with features obtained from the excitation source based on Linear Prediction (LP) residual and Hilbert envelop, when the speaker calls robot's name in all directions. For performance analysis, we collected SL-DB (sound localization database) with the variation of distance and angle under test-bed environments like home. Here the localization success rate (LSR) and average localization error (ALE) from field of view (FOV) range of robot camera are used as localization performance criterion. The experimental results reveal that the presented method shows a good performance in comparison with the well-known Time Delay of Arrival (TDOA) and Generalized Cross Correlation-Phase Transform (GCC-PHAT) method.

Keywords: Sound localization, excitation source information, intelligent home service robots, effective angle estimation, low-cost microphones.

1 Introduction

The recent years have been witnessed a considerable number of studies on Sound Localization (SL) for Human-Robot Interaction (HRI) under intelligent robot environments. Based on this technique, the robot can move and help for giving aid to a person by recognizing and judging a situation in public places and home. The most representative techniques frequently used in conjunction with SL method are Time Delay of Arrival (TDOA) [5] and Generalized Cross Correlation-Phase Transform (GCC-PHAT) [4][7]. The TDOA are widely used due to accuracy and simple computation. However, this method usually represents a poor localization performance under noise and reverberation environments. On the other hand, the GCC-based function is made more robust by deemphasizing the frequency-dependent weighting. However, the disadvantage of the PHAT weighting is that it places equal emphasis on both low and high SNR regions [1]. Various methods have been suggested for localization of speaker by modeling the production of speech. Most of the speech model-based methods use spectral features which correspond to the characteristics of the vocal tract system during the production of speech. The spectral features are affected by transmission through noise and room reverberation.

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 536 – 543, 2008.

Recently, a few attempts have been performed with the aid of the characteristics of the excitation source during the production of speech in conjunction with speaker recognition [2] and speaker localization [1]. In this study, we present and discuss on SL method using Excitation Source Information (ESI) and effective angle estimation for intelligent home service robots equipped with multi-channel sound board called MIM(Multimodal Interface Module) in ETRI(Electronics and Telecommunications Research Institute) and low-cost condenser microphones. In contrast to typical SL method in robot environments, the underlying principle exploited here is to consider a three-phase development of robust SL method. First, we use Endpoint Detection (EPD) algorithm based on log energy on SL-DB (sound localization database) obtained from three microphones equipped in intelligent mobile robots [6]. Here the database is collected by the variation of distance and angle under test-bed environments like home. Next, we compute a time-delay between two speech signals using ESI on SL-DB. Finally, we estimate a reliable localization angle from several candidate angles obtained by time-delay estimation of ESI. This paper is organized as follows. Section 2 describes the previous works on the well-known TDOA and GCC-PHAT frequently used in conjunction with SL method. In Section 3, we present a method for estimating time-delay using ESI. In Section 4, the SL-DB and the SL method based on ESI and effective angle estimation are described. Based on SL-DB, we perform comprehensive experiments in Section 5. Finally, conclusions are given in Section 6.

2 Previous Works: TDOA and GCC-PHAT

In this section, we describe the well-know previous works frequently used in conjunction with SL method such as TDOA and GCC-PHAT. These methods are compared with the presented technique on the constructed SL-DB in Section 5. Firstly we explain a brief review on time-delay estimation of TDOA. Let's consider windowed frames of N samples with 50% overlap. We assume that the index corresponding to each frame is omitted from the equations. It is necessary to define a coherence measure to determine time delay with the signals captured by two different microphones. The most well-known coherence measure is a simple cross-correlation between the signals perceived by two microphones as the following expression

$$R_{ij}(\tau) = \sum_{n=0}^{N-1} x_i[n]x_j[n-\tau] \tag{1}$$

where $x_i[n]$ is the signal received by i'th microphone and τ is the correlation lag in samples. The cross-correlation is maximal when τ is equal to the offset between the two received signals. However, this method is difficult to obtain a good estimate of the time-delay even when the signals are corrupted by noise and environments. On the other hand, GCC (Generalized Cross Correlation) is a correlation method in frequency domain. SL method based on GCC-PHAT has a lot of merits in noise environments and reverberation environments. When the signals $x_1(n)$ and $x_2(n)$ are

obtained by each of two microphones, the generalized cross-correlation between $x_1(n)$ and $x_2(n)$ can be obtained by the following equation.

$$R_{x_1x_2}(n) = \frac{1}{2\pi}\int_{-\infty}^{\infty} W(\omega)X_1(\omega)X_2^*(\omega)e^{j\omega n}d\omega \tag{2}$$

where $W(\omega)$ is a frequency weighting function. Among GCC-based methods, the most commonly used weighting function is PHAT (phase transform). Its weighting function is the reciprocal of $X_1(\omega)X_2^*(\omega)$. PHAT is a weighting function that determines the relative importance of each frequency as follows

$$W(\omega) = \frac{1}{\left|X_1(\omega)X_2^*(\omega)\right|} \tag{3}$$

The delay time between $x_1(n)$ and $x_2(n)$ can be obtained by the following expression

$$\tau = \arg\max R_{x_1x_2}(n) \tag{4}$$

3 Time-Delay Estimation of Excitation Source Information

In Linear Predication (LP) analysis, the sample $s(n)$ is estimated as a linear weighted sum of the past p samples. The predicted sample $\hat{s}(n)$ is given by

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{5}$$

where p is the LP's order and $a_k, k = 1,2,\ldots,p$, is LP's coefficients. These coefficients are obtained by minimizing the mean squared error (MSE) between the predicted sample value and the actual sample value over the analysis frame (40ms). The error between the actual value $s(n)$ and the predicted value $\hat{s}(n)$ is computed as follows

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{6}$$

This error $r(n)$ is called the LP residual of the speech signal. The LP residual contains information about the excitation source. The values of LP residuals are large around the instants of glottal closure for voiced speech. It is difficult to derive information from short segments of LP residual due to large fluctuations in amplitude. Here the analytic signal $r_a(n)$ corresponding to $r(n)$ is given by

$$r_a(n) = r(n) + jr_h(n) \tag{7}$$

The strength of the LP residual at each instant is obtained by computing the Hilbert envelope of the LP residual signal. The Hilbert envelope of the residual signal is obtained by the following expression [1][3]

$$h_e(n) = \left|r_a(n)\right| = \sqrt{r^2(n) + r_h^2(n)} \tag{8}$$

Fig. 1 shows the results of EPD based on log energy. Here the speech signal of robot name used in this study is wever. Fig. 2 visualizes the results obtained by LP residual and Hilbert envelope from two signals obtained through two microphones. Here we used only 250 ms (4000 samples) among the detected signal to perform fast computation. Fig. 3 shows the estimation results of the time-delay using ESI.

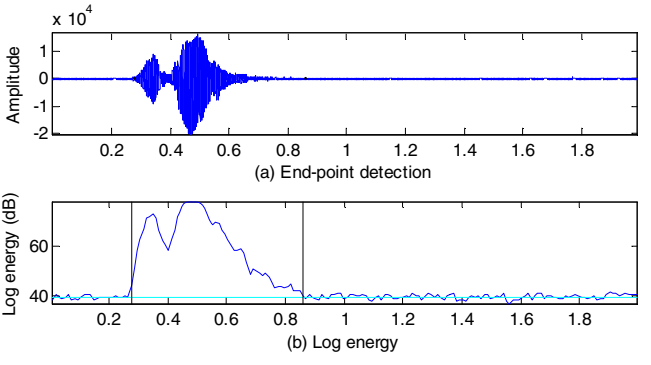

Fig. 1. Speech signal detected by EPD

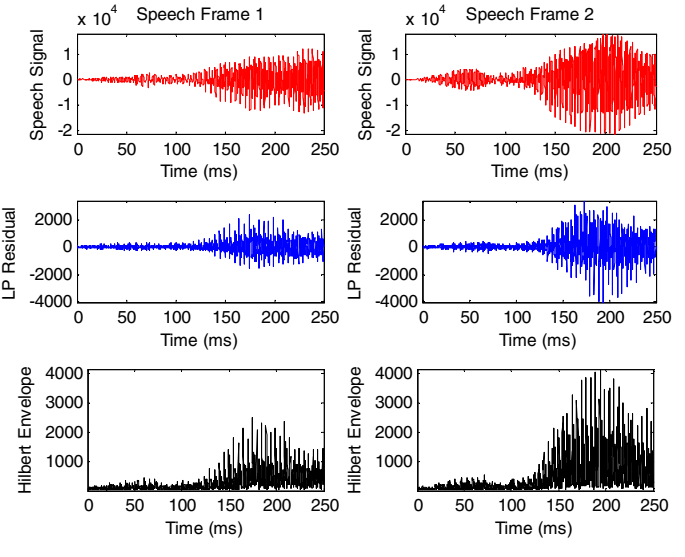

Fig. 2. LP residual and Hilbert envelope

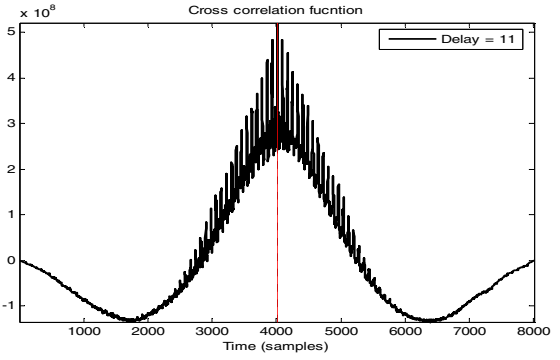

Fig. 3. Time-delay estimation by ESI

After computing time-delay, we obtain the localization angle between two microphones. Suppose that the sound wave at the microphone is a plan wave and the angle between microphones is 120°, respectively. Finally azimuth θ is estimated as follows

$$\theta = \cos^{-1}\left(\frac{\Delta t\, v}{L}\right) - 30 \tag{9}$$

where θ is the angle of sound source obtained in the above equation. Δt is time delay between two microphones and v is the velocity of sound source. Moreover L is the distance between two microphones.

4 Sound Localization Under Robot Environments

In this section, we describe SL-DB collected in test-bed environments like home. Furthermore, in order to localize sound source under robot environments, we present the method to estimate the effective localization angle from three time-delay values obtained by ESI.

4.1 SL-DB and Robot Environments

The SL-DB used in this study was constructed in test-bed environment that is similar with home environment to evaluate the SL algorithm. We used robot's name (wever) to collect SL-DB. The two individuals speak three times at 45 ° interval from 0 ° to 360 ° . The data set consists of 72 speeches at each meter from 1meter to 3 meter (M1 and M2 DB) when the number of low-cost condenser microphone is three. The audio is stored as a mono, 16bit, and 16kHz. Firstly we obtain the segmented speech signals from the database based on EPD algorithm. Fig. 4 shows the arrangement of three microphones with 120 degrees interval on Wever robot. The three microphones are equipped with multi-channel sound source board that is developed by ETRI. These microphones consist of low-cost condenser about four dollars, while the microphones presented in the previous literature are high-priced microphones.

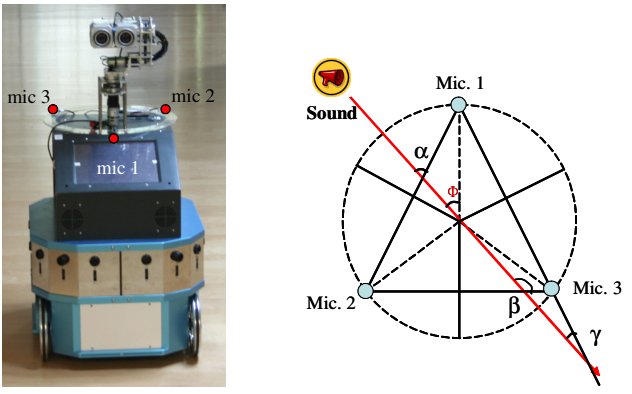

Fig. 4. Wever robot and the arrangement of microphones

4.2 Effective Angle Estimation

In what follows, we propose the method to estimate reliable localization angle from several candidate angles obtained by ESI. The three angles are obtained based on each time-delay obtained by ESI as the following equations

$$\alpha = \cos^{-1}\left(\frac{\mathrm{v}\tau_{12}}{\mathrm{d}}\right) \tag{10}$$

$$\beta = \cos^{-1}\left(\frac{\mathrm{v}\tau_{23}}{\mathrm{d}}\right) \tag{11}$$

$$\gamma = \cos^{-1}\left(\frac{\mathrm{v}\tau_{13}}{\mathrm{d}}\right) \tag{12}$$

where d (32cm) is a distance between each microphone and v (347.1 m/sec^2) is the velocity of sound, τ_{12} is time delay between channel 1 and channel 2, τ_{23} is time delay between channel 2 and channel 3, τ_{13} is time delay between channel 1 and channel 3. From these angles, we obtain six candidate angles as follows

$$\Phi_1 = \alpha - 30\ ^\circ \Phi_2 = -\alpha - 30\ ^\circ \Phi_3 = \beta + 90\ ^\circ \tag{13}$$
$$\Phi_4 = -\beta + 90\,, \Phi_5 = \gamma + 30\,, \Phi_6 = -\gamma + 30$$

Here because it is difficult to obtain ideal time-delay, we have to select the two closest Φ. And then we estimate the reliable localization angle by averaging these two angles.

5 Experimental Results

In this section, the presented approach is compared with TDOA and GCC-PHAT on SL-DB. The localization success rate (LSR) and average localization error (ALE) are

considered as performance measure. The LSR is computed by FOV(±15) of robot camera because SL method is used with face detection when robot moves toward caller. The experimental results are listed in Table 1. As summarized in table 1, the results of both LSR and ALE for M1 and M2 DB revealed that the presented method showed a better localization performance (more than 20%) in comparison to that of TDOA and GCC-PHAT. Here we only used 250 ms (4000 samples) period among whole speech signal detected by EPD algorithm. Table 2 lists six candidate angles and final estimated angle of the case of 0 degree in M1 data set.

Table 1. Performance comparison for M1 and M2 DB (during 250 ms, FOV±15)

	M1 set		M2 set	
	LSR (%)	ALE (degree)	LSR (%)	ALE (degree)
TDOA	79.2	5.44	63.9	5.29
GCC	56.9	5.97	81.9	4.53
The presented method	97.2	4.43	87.5	4.66

Table 2. Six candidate angles and final estimated angle (0 degree, 1~3 m)

	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6	final angle
1m	11.77	-71.77	176.11	3.88	65.55	-5.55	7.83
	11.77	-71.77	176.11	3.88	65.55	-5.55	7.83
	11.77	-71.77	-176.11	-3.88	77.31	-17.31	-10.60
2m	11.77	-71.77	176.11	3.88	77.31	-17.31	7.83
	-1.80	-58.19	180.00	0	58.19	1.80	-0.90
	5.55	-65.55	180.00	0	65.55	-5.55	2.77
3m	-11.64	-48.35	-172.20	-7.79	48.35	11.64	-9.71
	11.77	-71.77	-176.11	-3.88	77.31	-17.31	-10.60
	-11.64	-48.35	180.00	0	58.19	1.80	0.90

6 Conclusions

We have developed SL method with the aid of ESI and effective angle estimation for intelligent home service robots. The experimental results regarding SL-DB used in this study revealed that the presented method showed a better localization performance in comparison to the TDOA and GCC-PHAT. As a result, we can naturally communicates face to face with home service robot through spontaneous speech

recognition with continuous words to provide useful information such as weather information and daily life schedule at a short distance. Furthermore, when somebody calls the robot's name at a long distance, home service robot can detect the direction of sound source in all directions and then move forward caller.

References

1. Raykar, V.C., Yegnanarayana, B., Prasanna, S.R.M., Duraiswami, R.: Speaker localization using excitation source information in speech. IEEE Trans. on Speech and Audio Processing 13(5), 751–761 (2005)
2. Murty, K.S.R., Yegnanarayana, B.: Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Processing Letters 13(1), 52–55 (2006)
3. Rao, K.S., Prasanna, S.R.M., Yegnanarayana, B.: Determination of instants of significant excitation in speech using Hilbert envelope and group delay function 14(10), 762-765 (2007)
4. Knapp, C.H., Carter, G.C.: The generalized correlation method for estimation of time delay. IEEE Trans. on Acoustic, Speech, and Signal Processing ASSP-24, 320–327 (1976)
5. Huang, J., Supaongprapa, T., Terakura, I., Wang, F., Ohnishi, N., Sugie, N.: A model based sound localization system and its application to robot navigation. Robotics and Autonomous Systems, 199-209 (1999)
6. Kwak, K.C., Kim, H.J., Bae, K.S., Yoon, H.S.: Speaker identification and verification for intelligent service robots. In: Int. Conference on Artificial Intelligence (ICAI 2007), Las vegas, pp. 515–519 (2007)
7. Park, B.C., Ban, K.D., Kwak, K.C., Yoon, H.S.: Sound source localization based on audio-visual information for intelligent service robots. In: The 8th Int. Symposium on Advanced Intelligent Systems (ISIS, Sokcho 2007), pp. 364–367 (2007)