A SOUND-SOURCE LOCALIZATION SYSTEM USING THREE-MICROPHONE ARRAY AND CROSSPOWER SPECTRUM PHASE

HUNG-YAN GU, SHAN-SIANG YANG

Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology
Taipei 106, Taiwan
E-MAIL: guhy@mail.ntust.edu.tw, m9515043@mail.ntust.edu.tw

Abstract:

A sound source localization system is implemented that uses only three microphones to input sound signals. This system can estimate the azimuth and elevation of a sound source in real-time and in sufficient accuracy. We add a SNR measure besides spectra entropy to help detect voiced frames. Next, synchronous FFT phase copying is adopted, and cross-power spectrum phase is calculated to estimate TDOA (time delay of arrival) for each frame. Also, to enhance the accuracy of TDOA, parabolic interpolation is adopted. Then, by comparing the estimated TDOA values with theoretic ones, the azimuth and elevation of a sound source can be determined. Since a pair of azimuth and elevation is estimated from each voiced frame, these estimated values are thereafter summed with a weighting method to give one final answer of azimuth and elevation. According to the experiment results, the average errors in estimating azimuth and elevation are 4.02 and 2.18 degrees, respectively.

Keywords:

sound source localization; direction estimation; microphone array; TDOA; crosspower spectrum phase

1. Introduction

The technique of sound-source localization can be applied to interactive toys, robots, *etc*. If a humanoid robot can detect the direction of a person through his voice, the interactivity between them will be improved a lot. Here, direction is defined as a vector of azimuth (denoted with θ) and elevation (denoted with ϕ), i.e. $<\theta, \phi>$. Currently, there are two approaches to study the problem of direction estimation for a sound-source. In one of the approaches, the sound signals collected from a microphone array are used to compute a correlation matrix first. Next, the technique of beam-forming or subspace-theory [1] is used to estimate the direction of a sound-source. In another approach, the time delay of arrival (TDOA) between each pair of microphones is estimated first. Then, the geometric relationship between the sound-source and the microphone array is utilized to estimate the direction of the sound-source [2, 3].

Because the number of required computations is very large for the first approach, we decide to study with the second approach, i.e. TDOA based. Besides trying to improve the processing steps for direction estimation, we have practically built a real-time system to estimate the direction of a sound-source. The structure of our system is shown in Figure 1. It includes the following components:

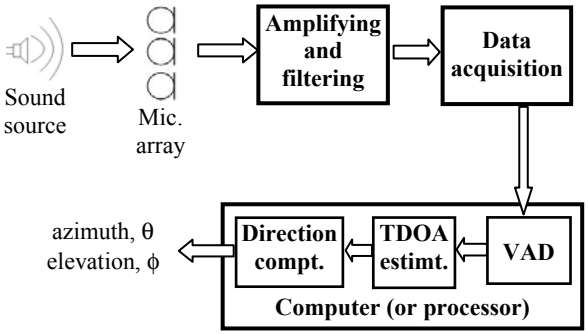

Figure 1. The structure of the proposed sound-source localization system.

- *Microphone array*: We use three microphones to form a right-triangular planar array. The distance between each pair of microphones is 20 cm. The orientation of the triangle is as indicated by the symbol, ∇.
- *Signal amplifying and filtering*: This circuit is used to amplify the signal sensed by each microphone and to low-pass filter the amplified signal. The cutoff frequency is set to 7,000 Hz.
- *Data acquisition*: This circuit first converts the analog signal from each microphone to its corresponding digital signal under the sampling rate 16,000 Hz. Then, the digitized signals are sent to the computer (or processor) through a USB line.
- *Voice Activity Detection* (VAD): This module computes the spectral entropy and SNR of each signal frame from each microphone. Then, it judges whether an input frame is a voice or noise frame according to

the entropy and SNR values computed from this frame.

- *TDOA estimation*: This module computes a generalized cross-correlation (GCC) function for each pair of microphones. In terms of each GCC function, a TDOA value is then estimated.
- *Direction computation*: In this module, three estimated TDOA values for a voiced frame are compared with the pre-computed theoretic values to determine a direction vector. Then, the direction vectors obtained from several frames are summed with a weighting method to give a global answer.

2. Voice Activity Detection

Three streams of sound signals from the microphones are continuously digitized by the data acquisition module, and brought to the computer box in Figure 1. Therefore, it is necessary to do voice activity detection in order to keep only those signal frames that contain voice signals. Here, three parallel frames from the three streams must all be detected as voice frames to ensure these frames contain voice signals. After detected as containing voice signal, each of the three parallel frames will be further processed to estimate a TDOA value. For VAD, we proposed a method that is based on spectral entropy and is enhanced with SNR verification.

2.1. Spectral Entropy

In the study by Renevey and Drygajlo [4], the probability of a frequency band is defined as the power of this band divided by the total power of the spectrum. Let $P(w, t)$ denote the probability of the w-th frequency band in the spectrum of the t-th signal frame. That is, $P(w, t)$ is defined as

$$P(w,t)=\frac{|Y(w,t)|^2}{\sum_{w=6}^{256\text{-}1}|Y(w,t)|^2},\quad w=6,\ 7,\ \cdots,\ 256\text{-}1\ , \tag{1}$$

where $|Y(w, t)|$ represents the magnitude of the w-th frequency band, and the number, 256, is a half of the frame size, 512. Here, 512 points FFT is executed to compute a signal frame's spectrum. In terms of $P(w, t)$, the spectral entropy of the t-th frame, H_t, can thus be calculated as

$$H_t=-\sum_{w=6}^{255}P(w,t)\cdot\log(P(w,t)) \tag{2}$$

When (2) is used, it is found in practice that the calculated entropy values for noise frames are not flat and may vibrate randomly. Hence, a solution method is proposed by Renevey and Drygajlo [4]. The method is to mix a frame with small white noise signal before that frame is taken to calculate entropy. Nevertheless, we find that the entropy values for a sequence of noise frames still vibrate significantly. Hence, we propose a further improving method, which is to filter the entropy curve with the formula,

$$\bar{H}_t=max\{H_t,H_{t-1}\} \tag{3}$$

2.2. SNR Verification

According to the entropy values calculated, a threshold can be set up to classify a frame into a voice frame or noise frame. Nevertheless, some voice frames may be incorrectly classified to, i.e. they are actually noise frames. This is due to a higher threshold value is selected to insure that true voice frames will almost be classified correctly. Therefore, it is needed to verify those frames that are classified as voice frames.

Note that those noise frames incorrectly classified as voice frames have smaller energies than the energy of a true voice frame. Therefore, our verification method is to compare the energy of a frame classified as voice frame with the average energy computed from those frames classified as noise frames. Actually, the comparison is to calculate an SNR,

$$SNR=10\cdot\log_{10}\frac{S}{N} \tag{4}$$

where S denotes the energy of the frame to be verified, and N denotes the average energy of the recently detected noise frames. According to the results of preliminary experiments, we decide to set the SNR threshold to 12 dB. That is, a frame of SNR greater 12 will pass the verification here. Otherwise, it will be reclassified as a noise frame.

3. Estimation of TDOA

Knapp and Carter proposed a generalized cross correlation (GCC) based method for estimating time delay [5]. One common definition of GCC is

$$R_{y_1y_2}(\tau)=\int_{-\infty}^{\infty}\frac{G_{x_1x_2}(f)}{\left|G_{x_1x_2}(f)\right|}\cdot e^{j2\pi f\tau}df\quad , \tag{5}$$

where $G_{x_1x_2}(f)$ denotes the cross power spectrum between two input signals, x_1 and x_2. In addition, $G_{x_1x_2}(f)$ is defined as

$$G_{x_1x_2}(f)=\int_{-\infty}^{\infty}R_{x_1x_2}(\tau)\cdot e^{-j2\pi f\tau}d\tau\quad , \tag{6}$$

$$R_{x_1x_2}(\tau)=\mathrm{E}\left[x_1(t)\cdot x_2(t-\tau)\right]\quad , \tag{7}$$

In (7), E[·] denotes the operation of taking expectation. Theoretically, a value of τ that let $R_{y_1y_2}(\tau)$ in (5) reach a maximum is the right time delay to be found.

For the consideration of practical implementation, we will not compute $G_{x_1x_2}(f)$ in terms of (6) and (7). Instead, according to the study by Omologo and Svaizer [6], we will first compute DFT spectrums, $X_1(f)$ and $X_1(f)$, from two signal frames of x_1 and x_2, respectively. Next, $X_1(f)$ and $X_1(f)$ are used to approximate $G_{x_1x_2}(f)$ and $|G_{x_1x_2}(f)|$ in (5). In details, we use $X_1(f)\cdot X_2^*(f)$ to approximate $G_{x_1x_2}(f)$, and use $|X_1(f)\cdot X_2^*(f)|$ to approximate $|G_{x_1x_2}(f)|$.

Additionally, we replace the integration in (5) with an inverse DFT. Due to the division of $X_1(f)\cdot X_2^*(f)$ by its magnitude, $|X_1(f)\cdot X_2^*(f)|$, the result of the inverse DFT (IDFT) will be a discrete pulse function, $\delta(k)$, $k = 0, 1, \ldots, 511$. Then, the delay time in discrete samples can be estimated as the value of the index, k, that maximizes $\delta(k)$. In practice, $\delta(k)$ is not an ideal pulse function but a rough one because of the approximation with the DFT spectrums.

As a phenomenon of speech signal, the magnitudes of $G_{x_1x_2}(f)$ in higher frequencies are usually much smaller than those in lower frequencies, which result in the phase values of $G_{x_1x_2}(f)$ in higher frequencies become unstable. Therefore, we have studied a method to solve this problem. That is, replacing the phases of higher frequency bins with those of their corresponding lower frequency bins in a synchronous manner. Precisely, we do phase replacing according to the formula,

$$\overline{\theta}_p(k) = 2\cdot\theta_p(\tfrac{k}{2}),\ k = 128,\cdots,255, \tag{8}$$

where $\theta_p(k)$ denotes the phase of the k-th frequency bin.

Due to the use of DFT and IDFT, delay time estimation is originally counted in sampling points. To improve the accuracy, we study a quadratic interpolation based method. First, the value of the index, k, that maximizes $\delta(k)$ is searched between 0 and τ_m. Here, τ_m is calculated as 0.2 / 346 * 16,000 = 9.25, where 0.2 (20 cm) is the distance between two microphones, 346 (m/sec) is the propagation speed of a sound wave, and 16,000 is the sampling rates. Suppose the discrete delay time is found to be τ sample points. Then, we use the three pairs, $(\tau\text{-}1,\ \delta(\tau\text{-}1))$, $(\tau,\ \delta(\tau))$ and $(\tau\text{+}1,\ \delta(\tau\text{+}1))$, to interpolate a parabolic curve, $b(x)$. On $b(x)$, the value of x that maximizes $b(x)$ is calculated. Suppose the value calculated is x_c. Then, we will take x_c instead of τ as the more accurately estimated value of TDOA.

4. Direction Estimation

A few studies that apply the TDOA values to estimate the direction of a sound source had been published. Here, according to our microphone array configuration, we modify the method proposed by Rabikin, *et al*. [7], to estimate the direction of a sound source.

Let $<xm_1,\ ym_1,\ zm_1>$ and $<xm_2,\ ym_2,\ zm_2>$ be the coordinates of two microphones. Also, let $<xs_1,\ ys_1,\ zs_1>$ be the coordinate of a sound source. Then, the time spent by a sound wave to propagate to the two microphones will be

$$t_j = \frac{\sqrt{\left(xm_j - xs\right)^2 + \left(ym_j - ys\right)^2 + \left(zm_j - zs\right)^2}}{c},\quad j = 1, 2, \tag{9}$$

where c is the propagation speed set to 346 m/sec here. Accordingly, the theoretic TDOA value between the two microphones is $D_{(1,\,2)} = t_1 - t_2$.

According to our microphone-array configuration, three microphone pairs can be defined. Hence, for each predefined sound source location, three theoretic TDOA values, $D_{(1,\,2)}$, $D_{(1,\,3)}$ and $D_{(2,\,3)}$, can be computed for the three microphone pairs. Here, let DV denotes the vector, $<D_{(1,\,2)},\ D_{(1,\,3)},\ D_{(2,\,3)}>$. In our estimation method, we compute a DV for each predefined direction of a sound source beforehand, and collect these DV into an acoustic map (ACMP) [8]. As to the selection of predefined directions, we decide to divide both azimuth and elevation uniformly in 5 degrees. Note that the range of azimuth and elevation are both from -90 to 90 degrees. Therefore, in the ACMP, the number of DV vectors is totally (180/5 - 1) × (180/5 + 1) + 2.

When the direction of a sound source is to be estimated, the three TDOA values of the three microphone pairs are estimated first, and arranged into a vector, $EV = < E_1, E_2, E_3>$. Next, a similarity distance between EV and each DV in the ACMP is computed with a geometric distance measure. Then, the DV vector, DV_{min}, of the smallest similarity distance is picked out. Accordingly, the azimuth, AZ_{min}, and elevation, EL_{min}, that are originally based to calculate DV_{min} are outputted as the direction of the sound source.

In practice, an inputted voice command may be sliced into a sequence of voice frames, and a pair of azimuth and elevation will be estimated for each voice frame (precisely three parallel frames from the three signal streams). Therefore, we have studied a weighting method to combine the estimated azimuth and elevation pairs, $<\theta_k,\ \phi_k>$, k=1, 2, …, F (the number o frames), into a single pair $<\theta,\ \phi>$, and then report it as the final answer. The weighting method is as described by the formula,

$$<\theta,\ \phi> \ = \left(\sum_{k=0}^{F\text{-}1} w_k\right)^{\text{-}1} \times \sum_{k=0}^{F\text{-}1} w_k <\theta_k, \phi_k>, \tag{10}$$

where w_k denote the weight for the k-th voice frame. We have tried to take the short-time energy [9] and RMS (root mean square) waveform amplitude of a voice frame as the weight.

According the results of our experiments, the short-time energy is found to be the better choice to act as the weight.

5. Experimental Evaluation

Since a sound source may locate at various directions, it is necessary that same voice signals are uttered when the performance of our system is to be evaluated. Therefore, we first recorded 10 different voice commands uttered by two male and two female participants in a soundproof room. These recorded voice signals are then played back with a speaker placed at the directions to be tested. The distance between the speaker and the center of the microphone array is set to 100 and 150 cm, respectively. Here, the selected directions are the combinations of the seven azimuths (-90, -60, -30, 0, 30, 60, and 90) and the three elevations (-30, 0, 30). In each of the combined directions and distances, the 10 voice commands uttered by the four persons are played back one after one, and the direction of each voice command is estimated in real-time by our system.

For each voice command played back, a direction, $<\theta, \phi>$, is immediately estimated by our system and saved into a file. After conducting the experiments as mentioned above, we analyzed the saved direction data. When analyzed in respect of each person, the estimation errors in average (AVG) and standard deviation (STD) for azimuth (AZI) and elevation (ELE) are as those listed Table 1. When the numbers in Table 1 are averaged, it is obtained that the gross estimation error for azimuth is 4.02 degrees in average and 2.88 degrees in standard deviation, whereas for elevation the estimation error is 2.18 degrees in average and 1.93 degrees in standard deviation. In addition, it can be seen from Table 1 that there are no significant differences in AVG and STD error values between the male persons and the female persons.

TABLE 1. ESTIMATION ERRORS IN RESPECT OF EACH PERSON.

	Male A		Male B		Female A		Female B	
	AZI	ELE	AZI	ELE	AZI	ELE	AZI	ELE
AVG	3.46	2.04	4.89	2.71	3.65	1.89	4.09	2.10
STD	1.22	0.86	3.36	2.42	2.07	1.42	2.82	1.84

As another analysis manner, we analyze the estimation errors in respect of the azimuth and distance of the speaker that is placed at to play back the recorded voice commands. For different combinations of distances and azimuths, the estimation errors of azimuth in average and standard deviation are as those listed in Table 2. As for the estimation errors of elevation, they are listed in Table 3.

From Table 2 and Figure 2, it is seen that the estimation errors become larger especially between 60 (or -60) and 90 (or -90) degrees when the sound source is moved from the front to the left or right side of the microphone array in both tested distances. This is because the TDOA value is not linearly varied with the azimuth value. As to the worst estimation error, i.e. AVG 11.0 degrees, it is encountered when the speaker is placed at the azimuth of 90 degrees and the distance of 150 cm. This worst error value is attributed to the fact that a rotating fan that generates noise is very close to that location. In contrast, according to Table 3 and Figure 3, it seems that there are no significant differences in the error values of elevation estimation.

TABLE 2. AZIMUTH ESTIMATION ERRORS IN RESPECT OF SOUND SOURCE DISTANCE AND AZIMUTH.

Distance	Err.	AZI 90	AZI 60	AZI 30	AZI 0	AZI -30	AZI -60	AZI -90
100 cm	AVG	**5.72**	**2.56**	**3.38**	**2.11**	**2.14**	**2.81**	**5.41**
	STD	2.59	1.42	0.91	1.06	1.49	2.23	1.88
150 cm	AVG	**11.00**	**2.99**	**2.88**	**2.33**	**2.67**	**3.23**	**7.10**
	STD	6.64	3.06	1.58	1.21	1.88	3.06	4.12

TABLE 3. ELEVATION ESTIMATION ERRORS IN RESPECT OF SOUND SOURCE DISTANCE AND AZIMUTH.

Distance	Err.	AZI 90	AZI 60	AZI 30	AZI 0	AZI -30	AZI -60	AZI -90
100 cm	AVG	**1.91**	**1.19**	**0.97**	**2.34**	**1.96**	**2.35**	**3.23**
	STD	1.38	0.75	0.72	1.05	1.33	1.27	1.34
150 cm	AVG	**2.67**	**1.65**	**1.47**	**2.39**	**2.43**	**3.10**	**2.92**
	STD	3.40	1.99	1.06	1.51	1.99	2.82	2.23

As for computation efficiency, we used a notebook computer with an Intel Core 2 Duo T8300 CPU and 2 GB memory to run our system to estimate sound-source directions. According to the displaying of the resource monitor program provided by Windows XP OS, it is seen that the largest of the CPU rates displayed is 13% during the execution of our system. Therefore, our system is satisfactory for real-time sound-source localization applications.

6. Conclusions

In this paper, a sound-source localization system is implemented with a triangular three-microphone array and cross-power spectrum phase based TDOA estimation. This system can estimate the direction, i.e. azimuth and elevation, of a sound source in real-time, only 13% CPU time at most is consumed. According to the results of the direction estimation experiments conducted, it is shown that our system can achieve a satisfactory performance. The average errors in estimating azimuth and elevation are as low as 4.02 and 2.18 degrees, respectively.

For VAD, we have proposed a filtering method to reduce the vibration of the measured entropy curve. In addition, a SNR based verification method is designed to prevent a noise frame from being incorrectly classified as a voice frame. For the estimation of TDOA values, we propose to replace the phases of the higher frequency bins with those of their corresponding lower frequency bins in a synchronous manner. This helps to have stable TDOA sample points be estimated. Additionally, in terms of parabolic interpolation, the accuracy of the estimated TDOA values is significantly increased.

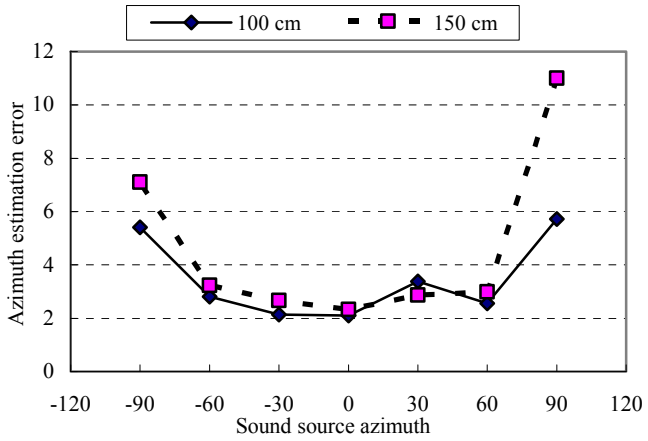

Figure 2. Azimuth estimation errors vs. sound source distances and azimuth.

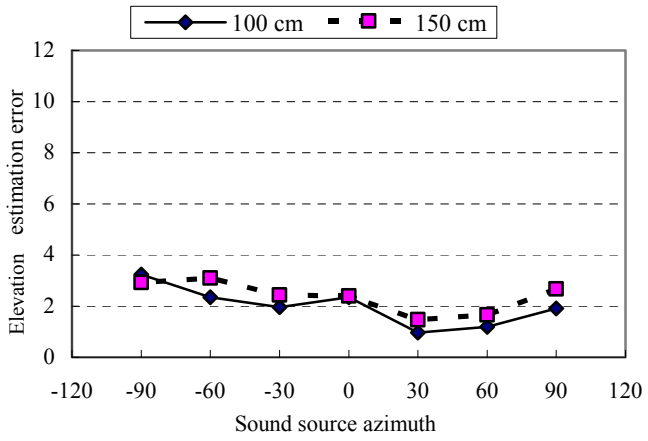

Figure 3. Elevation estimation errors vs. sound source distances and azimuth.

Acknowledgement

This study is sponsored by National Science Council, Taiwan, under the contract number, NSC 96-2218-E-011-002.

References

[1] R. O. Schmidt, “Multiple emitter location and signal parameter estimation”, IEEE trans. Antennas and Propagation, vol. AP-34, no.3, pp. 276-280, 1986.

[2] B. Kwon, G. Kim and Y. Park, “Sound source localization methods with considering of microphone placement in robot platform”, The 16th IEEE Int. Symposium on Robot and Human Interactive Communication, Jeju, Korea, pp. 127-130, 2007.

[3] X. Lv and M. Zhang, “Sound source localization based on robot hearing and vision”, Int. Conf. Computer Science and Information Technology, Singapore, pp. 942-946, 2008.

[4] P. Renevey and A. Drygajlo, “Entropy based voice activity detection in very noisy conditions”, 7th European Conference on Speech Communication and Technology (EuroSpeech), Aalborg, Denmark, 2001.

[5] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay”, IEEE trans. Acoustics, Speech and Signal Processing, Vol. 24, No. 4, pp. 320-327, 1976.

[6] M. Omologo and P. Svaizer, “Use of the crosspower spectrum phase in acoustic event location”, IEEE trans. Speech and Audio Processing, Vol. 5, No. 3, pp. 288-292, 1997.

[7] D. V. Rabikin, R. J. Renomeron, A. Dahl, J. C. French, and J. Flanagan, "A DSP implementation of source location using microphone arrays", in Proc. 131st Meeting of the Acoustical Society of America, Indianapolis, Indiana, pp. 88-99, 1996.

[8] A. Brutti, M. Omologo, and P. Svaizer, “Comparison between different sound source localization techniques based on a real data collection”, Hands-free Speech Communication and Microphone Arrays, Trento, Italy, pp. 69-72, 2008.

[9] D. O’Shaughnessy, Speech Communications: Human and Machine, IEEE Press, Piscataway, NJ, 2000.