Asia-Pacific Journal of Science and Technology: Volume: 22. Issue: 02. Article ID.: APST-22-02-03. Research Article

APST

Asia-Pacific Journal of Science and Technology

https://www.tci-thaijo.org/index.php/APST/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Blind 2D signal direction detection with limited sensor space using maximum likelihood estimation

Naruephorn Tengtrairat,[1,] * Phetcharat Parathai,[1] Wai Lok Woo[2]

[1] School of Computer Science, School of Software Engineering, Faculty of Science, Payap University, Chiang Mai, 50000, Thailand.
[2] School of Electrical and Electronic Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom.
*Corresponding author: n.tengtrairat@gmail.com

Abstract

Humans can automatically turn toward the origin of a sound that they hear through their both ears. This capability is crucial in daily life. A computer that can recognize the direction of the source of a sound would also be useful for many types of applications. This paper describes the development of blind two-dimensional (2D) signal direction recognition using two recording sensors with limited space between the sensors in a noise-free environment. The proposed 2D source direction recognition method is based on time-delay estimation using maximum likelihood estimation by forming a histogram of a power-weighted spectrum corresponding to attenuation and time-delay index. A histogram-boundary method is also proposed that relates to a distance of the two microphones. In addition, the optimal number of time-index bins for the histogram were investigated. Given the narrowness of the space between microphones (i.e., 3.2 centimeters), the proposed method can acceptably estimate the position of the sound source. In experimental testing on real audio sources, the proposed method has demonstrated a higher level of directional performance than an existing method.

Keywords: Sound source direction, Blind stereo estimation, Time – delay estimation.

1. Introduction

Humans can automatically turn toward the origin of a sound that they hear through their both ears. This capability is crucial in daily life. A computer that can recognize the direction of the sound source would have many applications. For example, CCTV cameras could be made to rotate in the direction of the sound source. Such computers could also be used to assist the hearing impaired. Moreover, a robot could turn toward a human speaker in order to more convincingly engage in conversation. In recent years, several SSD (sound source direction) recognition approaches have been developed. These can mainly be classified into two frameworks, i.e., active and passive [1]. The passive SSD approach can be naturally applied to real world problems, for example, a robot can walk toward someone by following their voice or calculate the direction of gunfire by the sound. In order to resemble human auditory capability without adding an extra sensor, most research in this area has been focused on this passive framework. Generally, n microphones are used to detect audio in order to estimate the direction of the signals, where $n = 2,4,\dots,N$ denotes the number of microphones. Some studies [2-7] introduced these types of methods for determining the direction of a signal. The most common technique used to achieve this is an extension of the generalized cross-correlation (GCC) method [3-6]. The GCC method requires more than two microphones and an appropriate space between the microphones (usually about 0.5 m). In the field of source separation, a powerful approach for estimating time delay is the degenerate unmixing estimation technique (DUET) [8]. This technique extracts the time lag and attenuation a difference between the corresponding T-F unit (time-frame unit) of the two microphone recordings. The DUET algorithm has been shown to yield highly accurate estimations of the underlying time-delay and attenuation features of audio sources when given two anechoic mixtures in the TF domain (time-frequency domain).

However, the performance of the TDE approach is related to the distance between the microphones, as the accuracy dramatically declines when the distance is short. Thus, the validity of this method is dependent on proper microphone position. Few researches have investigated using a binaural sensor for determining source direction with limited space between microphones. Therefore, we conducted this research in order to develop a blind 2D signal-direction detector with limited space between the sensors using maximum likelihood estimation.

This paper is organized as follows: section two presents a proposed for blind two-dimensional (2D) signal direction estimation; section three elucidates a resolution setting; section four describes experimental results and discussion with a series of performance comparisons with other method of estimating the azimuth angle; and section five consists of a summary and conclusion of this research.

2. Proposal for blind two-dimensional (2D) signal direction detection using limited-space stereo methods

Two microphones (i.e., Mic.1 and Mic.2) are aligned on a horizontal plane at a distance D apart in an anechoic environment. The signals received by Mic.1 and Mic.2 can be expressed as:

$$\left.\begin{aligned} m_1(t) &= a_1 s(t-\delta_1) \\ m_2(t) &= a_2 s(t-\delta_2) \end{aligned}\right\}, \tag{1}$$

where $m_j(t)$ is the Mic.j, j^{th} is the index of the microphone, a_j is an attenuation coefficients of the j^{th} microphone, and δ_j denotes the time delay associated with the path from the source to the j^{th} microphone. Using the two recorded signals, the horizontal direction of the sound source can be computed by calculating the cosine of the angle of the arrival signals $m_1(t)$ and $m_2(t)$:

$$cos(\theta) = \frac{\hat{\delta} v}{D}, \tag{2}$$

where θ denotes the angle of azimuth, $\hat{\delta}$ is the time lag between the two recorders, and v is the sound velocity in air.

To begin, one microphone is used as a reference point and the input signals in Eq. (1) are transformed into a TF representation using the short-time Fourier transform (STFT):

$$\left.\begin{aligned} M_1(\tau,\omega) &= S_j(\tau,\omega) \\ M_2(\tau,\omega) &= a\, e^{-i\omega\delta} S_j(\tau-\delta,\omega) \end{aligned}\right\} \tag{3}$$

The time-delay estimator $\hat{\delta}$ can be obtained using the maximum-likelihood (ML) estimator. The derivation is detailed in a paper by Yilmaz and Rickard [8]. The ML estimate $\hat{\delta}$ will be used to cluster the TF units into a number of groups corresponding to the arrival difference between the two microphones and a weighted average of the signal coefficient by the product magnitude of $\sum_{\tau,\omega}|M_1(\tau,\omega)M_2(\tau,\omega)|$. Hence, the time-delay estimation $\hat{\delta}$ can be expressed as:

$$\begin{aligned} \hat{\delta} &= \frac{\iint_{(\tau,\omega)} |M_1(\tau,\omega)M_2(\tau,\omega)|^p \omega^q Im\left[\frac{M_2(\tau,\omega)}{M_1(\tau,\omega)}\right] d\tau d\omega}{\iint_{(\tau,\omega)} |M_1(\tau,\omega)M_2(\tau,\omega)|^p \omega^q \, d\tau d\omega} \\ &= \frac{\sum_{\tau,\omega} |M_1(\tau,\omega)M_2(\tau,\omega)|^2 \omega^2 Im\left[\frac{M_2(\tau,\omega)}{M_1(\tau,\omega)}\right]}{\sum_{\tau,\omega} |M_1(\tau,\omega)M_2(\tau,\omega)|^2 \omega^2} \end{aligned} \tag{4}$$

where Im(z) corresponds to the imaginary part of z, $p = 2$, and $q = 2$. It is assumed that the sources satisfy the local stationarity of the time-frequency representation. This refers to the approximation of $S_j(\tau - \phi, \omega) \approx S_j(\tau, \omega)$ where ϕ is the maximum time-delay (shift) associated with the STFT with an appropriate window function. Thus, the term $M_2(\tau,\omega)/M_1(\tau,\omega)$ can be derived by applying a local stationarity assumption as:

$$\frac{M_2(\tau,\omega)}{M_1(\tau,\omega)} = \frac{a\, e^{-i\omega\delta} S_j(\tau-\delta,\omega)}{S_j(\tau,\omega)}$$

$$= \frac{a\, e^{-i\omega\delta} S_j(\tau,\omega)}{S_j(\tau,\omega)}$$

$$= a\, e^{-i\omega\delta} \quad (5)$$

Then, substituting Eq. (5) into Eq. (4), the $\hat{\delta}$ can be estimated by:

$$\hat{\delta} = \frac{\sum_{\tau,\omega} |M_1(\tau,\omega) M_2(\tau,\omega)|^2 \omega^2 Im\left[a\, e^{-i\omega\delta}\right]}{\sum_{\tau,\omega} |M_1(\tau,\omega) M_2(\tau,\omega)|^2 \omega^2} \quad (6)$$

Once the estimated time-delay $\hat{\delta}$ is obtained in the TF domain, which is in a unit of time-frame index, $\hat{\delta}$ is converted back to the time domain using $\hat{\delta}_s = \hat{\delta}/f_s$ where f_s denotes sampling frequency, and the unit is seconds. Next, Equation (2) is computed by substituting the term $\hat{\delta}$. The azimuth is finally determined by solving Equation (2).

2.1. Determine boundary of histogram

The angles on the front of the microphones are in a range [0, π] that depends on the estimated time delay $\hat{\delta}$ from the ML estimate. The $\hat{\delta}$ estimation based on the ML method is computed by constructing a power-weighted histogram. To cover the angular range and the number of the different estimate angles, the histogram resolution must be properly determined. This will allow for an adequate difference in time-frame index to be obtained. Based on the association of the time delay in TF and time domain, a maximum time index of $\delta_{\max}$ can be computed:

$$\delta_{max} = \delta_{s_{max}} \times fs$$

$$= \frac{D}{v} \times fs \quad (7)$$

where $\delta_{s_{\max}} = D/v$. For example, if the distance between two microphones is 0.032 m, the speed in the air v is approximately at 340 m/s, and f_s was set at 44,100 Hz., $\delta_{\max}$ would be 4.15. If the resolution margin is set to over $\delta_{\max}$, the azimuth will be a complex value. On the other hand, if the resolution margin less than $\delta_{\max}$, the angular range will not be reached [0, π].

3. Results and discussion

The recoding array used on-the-shelf directional microphones i.e. Sony ICD-UX300 with a frequency range from 40 Hz to 20 kHz, sampling rate at 44.1 kHz and bit rate at 32kbps. The built-in stereo microphones were 3.2 cm apart and covered by a reflector of seven circular slits on each side. The recorder was placed on a table at 0.75 m. height and the sound source was at a distance of 1.65 m from the center of the microphone array. The microphone array was fixed and the source positions were located at five different angles on the horizontal plane i.e. 30, 60, 90, 120, and 150. The sound sources were ten-second recordings of humans speaking. The total number of sources were 50 realizations with 10 realizations for each angle position. The histogram-resolution parameters were set at $\delta_{max} = \pm 4.2$. The recorded signals were transformed from time to TF representation using the STFT of the 1024-point Hanning window with 50% overlap. Examples of recorded signals in time and TF domains are shown in Figures 1 and 2, respectively.

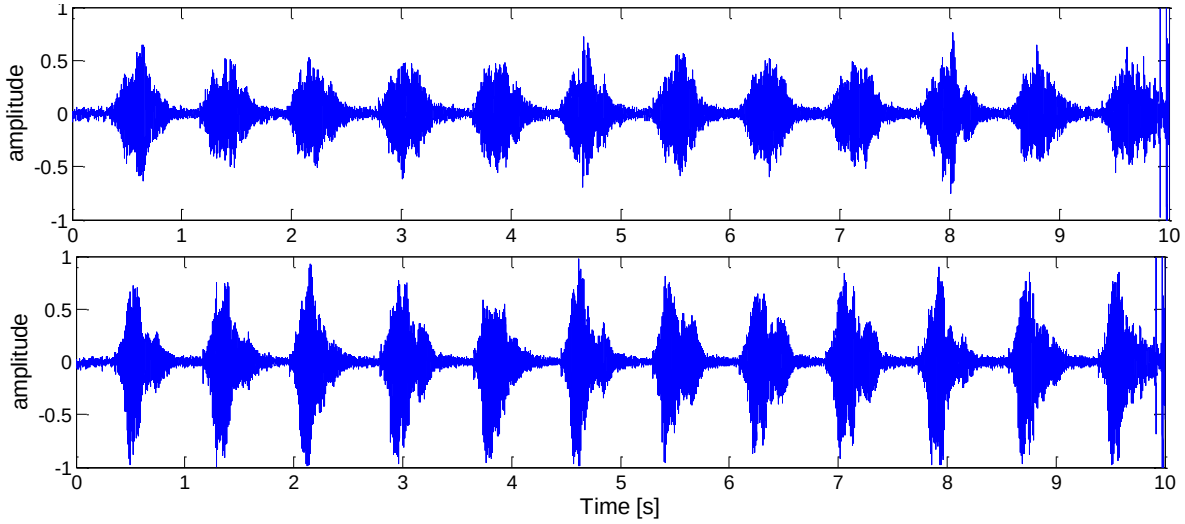

Figure 1 Recorded signals of left (above) and right (below) microphones at 30° in time domain.

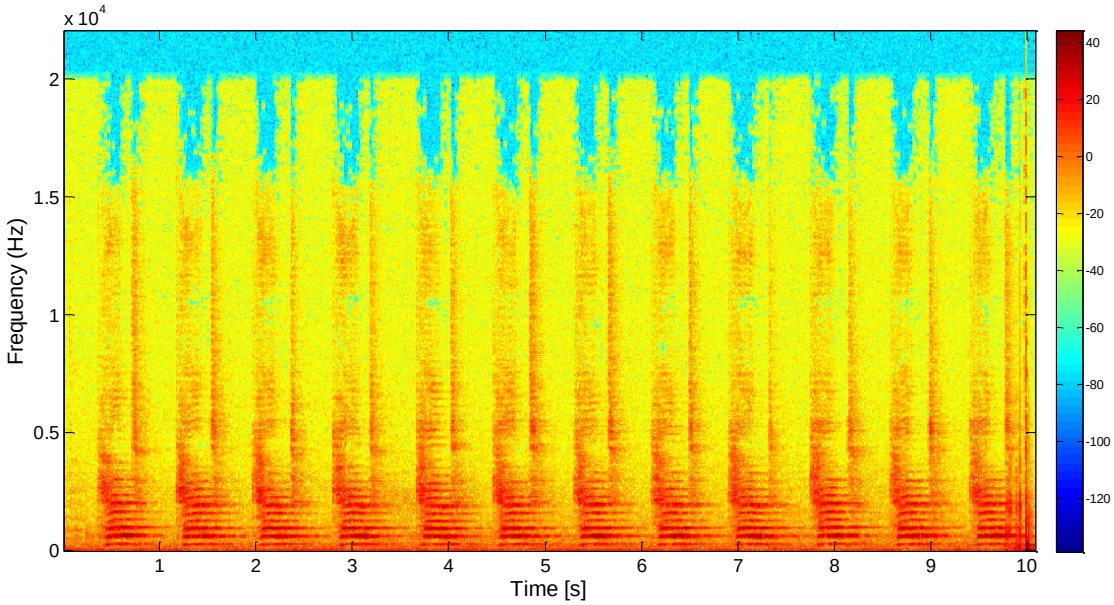

Figure 2 Recorded signals of left microphone at 30° in time-frequency domain.

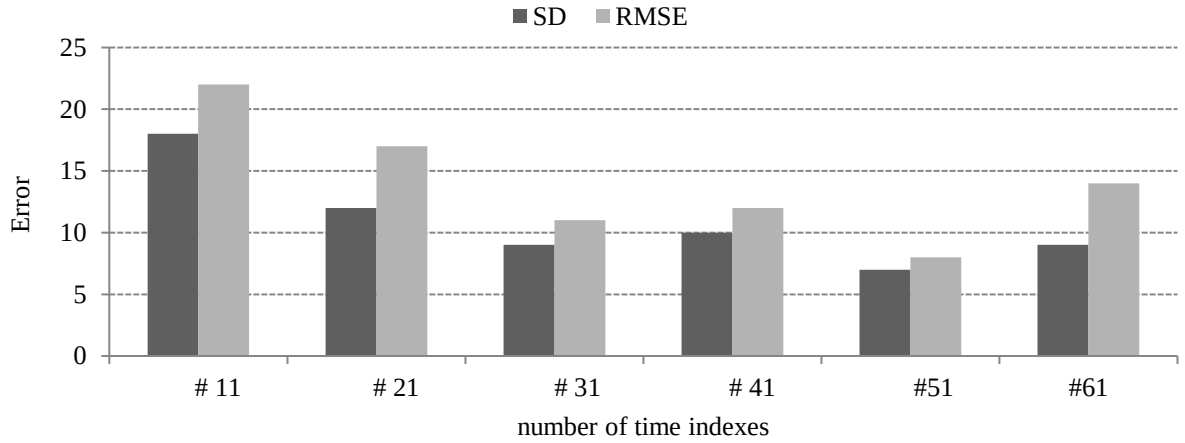

Figure 3 Comparison of average SD and RMSE of various time-index bins

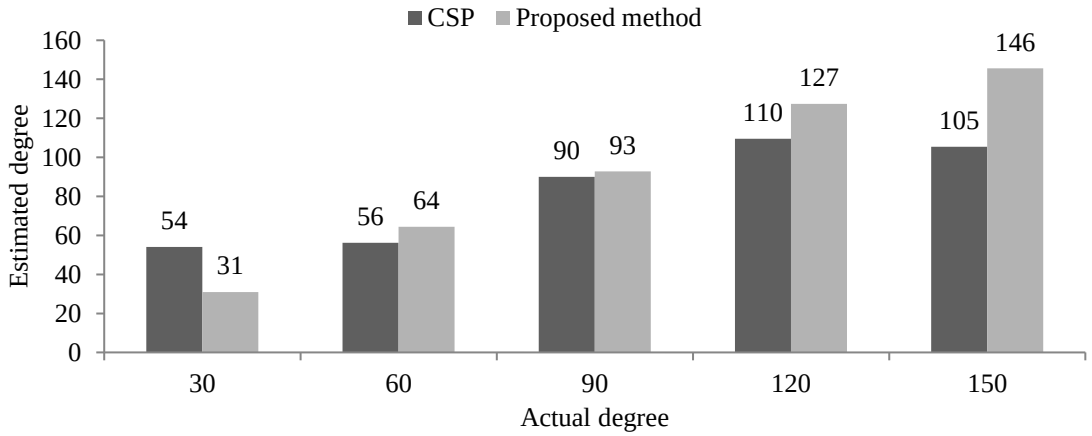

Figure 4 Comparison of average estimated angle of CSP and proposed method

3.1. Determination of time – delay bins in the histogram

The time-delay estimation was obtained from the selected point in the histogram. The time-index boundary of the histogram was set according to $\delta_{\max}$, where the resolution was based on the number of time-index bins (δ_b) between $\pm\delta_{\max}$. An experiment was conducted to determine the proper number of bins by examining the estimated performance given by various bins i.e. 11, 21, 31, 41, 51, and 61. The testing set for various bins was employed for all five different angle positions. The performance was evaluated using two metrics: standard deviation (SD) and root mean square error (RMSE). The results are shown in Figure 3.

When the histogram had 51 time-index bins, it yielded the smallest errors of average standard deviation (SD) and root mean square errors (RMSE) compared with the other numbers of bins. Therefore, the δ_b was set at 51 for all experiments.

3.2. Estimated angle results

The performance of the proposed method was demonstrated in a noise-free environment. The proposed method has been compared with the crosspower-spectrum phase (CSP) method [5]. The CSP method converts an input signal from time into the frequency domain and computes multiplications of the normalized spectra. To obtain a time-delay, the normalized crosspower-spectrum result is converted back into the time domain. The CSP parameter was set as 512-point Hanning window with 50% overlap.

The estimated angle performance of the proposed method has been assessed using 50 recorded signals and compared with CSP. The average estimated angles are presented in Figure 4.

Overall, the proposed method yields more accurate results, with a total average improvement of SD at 9 degrees and RMSE at 14 over the CSP method. This is because the CSP computes the time delay by weighting the complex cross spectrum to determine the phase relationship of the two recorded signals in a function of frequency. In cases of weak energy signals, the weight factor as a small denominator will cause an error with regard to the relationship between the two observed signals. Moreover, the limited distance between the sensors leads to a very short range of time delay between the microphones. Thus, the estimated delay of the proposed method is in a range of [-4.15, 4.15] with 51 bins to cover the angle from 0 to π. The time-delay estimation based on CSP is an integer value i.e. {-4, -3, …, 3, 4}. Thus, the angular resolution of CSP only has nine bins. This means that the proposed method ameliorates the problem of the narrow range of the delay by increasing the histogram resolution.

4. Conclusions

This paper proposed novel blind 2D sound source direction detection using 2 recording sensors in a small space. The time-delay estimation was based on the maximum-likelihood (ML) estimator, which was computed by constructing a power histogram. A method for determining the boundary of the histogram was also presented. The proposed method takes advantage of the adjustable histogram-resolution. Therefore, experimental results showed that the proposed method yields more reliable results than the CSP method.

5. Acknowledgements

This research was fully supported by Payap University. We are grateful to our consultant from the School of Electrical and Electronic Engineering at Newcastle University who provided insight and expertise that greatly

assisted the research. I would like to thank reviewers for their time, constructive criticism, and feedback. This paper is much better because of them.

6. References

[1] Khaddour, H., Schimmel, J., Trzos, M., 2013. Estimation of direction of arrival of multiple sound sources in 3d space using b-format. International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems 2, 63-67.

[2] Algazi, V.R., Avendano, C., Duda, R.O., 2001. Elevation localization and head-related transfer function analysis at low frequencies. The Journal of Acoustical Society of America 109, 1110-1122.

[3] Chen, L., Liu, Y., Kong, F., He, N., 2011. Acoustic source localization based on generalized crosscorrelation time-delay estimation. Procedia Engineering 15, 4912-4919.

[4] Lima, M.V., Martins, W.A., Nunes, L.O., Biscainho, L.W., Ferreira, T.N., Costa, M.V., Lee, B., 2015. Efficient Steered-Response Power Methods for Sound Source Localization Using Microphone Arrays. IEEE signal processing letters 22, 1-15.

[5] Marinescu, R.S., Buzo, A., Cucu, H., Burileanu, C., 2013. Applying the Accumulation of Cross-Power Spectrum Technique for Traditional Generalized Cross-Correlation Time Delay Estimation. International Journal On Advances in Telecommunications-IARIA 6, 98-108.

[6] Nakashima, H., Kawamoto, M., Mukai, T., 2010. A localization method for multiple sound sources by using coherence function. In: Signal Processing Conference, 2010 18th European. IEEE, 130-134.

[7] Pourmohammad, A., Ahadi, S.M., 2012. TDE-ILD-HRTF-Based 2D Whole-Plane Sound Source Localization Using Only Two Microphones and Source Counting. International Journal of Information and Electronics Engineering 2, 307-313.

[8] Yilmaz, Ö., Rickard, S., 2004. Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing 52, 1830-1847.