An Efficient Pre-processing Scheme to Improve the Sound Source Localization System in Noisy Environment

Sheng-Chieh Lee[1+], K. Bharanitharan[3], Bo-Wei Chen[1], Jhing-Fa Wang[1++], Chung-Hsien Wu[2], and Min-Jian Liao[1]

[1] Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan
[2] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
3Department of Electrical Engineering, Korea University, Seoul, Korea
[+]leesc@icwang.ee.ncku.edu.tw, [++]wangjf@mail.ncku.edu.tw

Abstract

In this study, we introduce an efficient pre-processing scheme for direction of arrival (DOA) estimation, which is capable of reducing the noise and reverberation effects in speech sound source localization. Furthermore, this presented system is also suitable for far-field speech localization. The adopted method of this proposed system can be simply subdivided into three stages: Linear phase-difference approximation, covariance matrix reconstruction, and frequency bin selection. The first two stages can initially decrease the influences of noise and reverberation; the last stage is used to filter the noise frequency bands according to the eigenvalue decomposition (EVD) of the covariance matrix. The experimental results show that our proposed system has effective performance of detecting different directions of speeches. For different signal-to-noise ratios (SNRs) speech signals, the average estimation errors can be decreased by about 5 to 7.5 degrees.

Index Terms: sound source localization, direction of arrival, phase-difference approximation, frequency bin selection

1. Introduction

The objective of sound source localization (SSL) is applied to determine the voice location coordinates in a restrictive space. More recently, research on SSL has been explored in number of scenarios, including audio-visual conferencing, surveillance, and speech recognition. Since SSL has been an active research topic in acoustic and speech signal processing, many studies are presented to deal with this demand for each different SSL situation.

In general, the SSL system is composed of a microphone array and an adopted direction of arrival (DOA) algorithm. The former is used to receive the sound source, and the latter is utilized to evaluate the position of the received sound source. The DOA estimation algorithms can be ordinarily divided into two categories, the first one is time delay estimation (TDE) method, and the other one is eigen space distribution approach.

The TDE method commonly employs a receiver array, which consists of only two sensors, to obtain sound signals. Each signal passing from different direction propagation in the sensor array causes various responses. The TDE algorithms, such as average magnitude difference function (AMDF) [1] and generalized cross correlation (GCC) [2], can calculate the direction angle of signal source according to the time difference of paths, where the signal transmits from source to the diverse sensors. On the other hand, the eigen space-based approaches like multiple signal classification (MUSIC) [3] can compute the multiple directions of the signals with the relation of eigenvector distributions between different signals and the orthogonal projections of individual eigenvectors.

When the sound source belongs to narrowband signal, the accurate DOA estimation can be achieved by the conventional DOA algorithms mentioned above. However, the performance of wideband signal (e.g., speech signal) DOA evaluation with the usual narrowband DOA algorithms could be extremely decreased in ambient noise and reverberation environments. It is because noise may have the same spectral characteristics as the desired speech signal, and reverberation causes an increment in the phase difference of received signals between microphones. In order to handle this problem, some of studies have been proposed the spatial smoothing technique [4] and the distributed microphone system [5]. However, it is hard to implement while the number of microphones is restricted. For the reasons described above, the purpose of this study is to investigate a pre-processing procedure of DOA estimation with only two microphones, which can relieve the noise and reverberation problems in a real environment.

The remainder of this study is organized as follows: Section II reviews the previous work of DOA algorithms. In Section III, we present the details of the proposed DOA system with pre-processing scheme. The experimental results are examined in Section IV. Section V briefly summarizes the conclusions of this study.

2. Previous work

2.1. Multiple signal classification

Multiple signal classification (MUSIC) is a subspace-based approach for DOA estimation that utilizes the EVD to separate the noise space and the desired space from the received signal, and to calculate the direction of the source signal by the eigenvector of the covariance matrix.

We assume that the source signal is collected by the microphone array, which is composed of m microphones; the received signals can be modeled with a matrix or vector terms, where $a(\theta)$ denotes a steering vector, $s(t)$ and $n(t)$ represent the sound signal and the noise respectively.

$$\begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_m(t) \end{bmatrix} = \begin{bmatrix} a_1(\theta)s(t) + n_1(t) \\ a_2(\theta)s(t) + n_2(t) \\ \vdots \\ a_m(\theta)s(t) + n_m(t) \end{bmatrix} \quad (1)$$

$$\mathbf{u}(t) = \mathbf{A}(\theta)s(t) + \mathbf{n}(t) \quad (2)$$

Hence, the covariance matrix of the observed signal can be given in (3), where $E[\bullet]$ and $\bullet^H$ are the expectation value and the Hermitian operation, $\mathbf{R}_{ss}$ means the covariance matrix of the sound source, σ_n^2 denotes the variance of the noise, and $\mathbf{I}$ is an identity matrix.

$$\begin{aligned}\mathbf{R_{uu}} &= E[\mathbf{u}\mathbf{u}^H] \\ &= \mathbf{A}E[\mathbf{s}\mathbf{s}^H]\mathbf{A}^H + E[\mathbf{n}\mathbf{n}^H] \\ &= \mathbf{A}\mathbf{R_{ss}}\mathbf{A}^H + \sigma_n^2\mathbf{I}\end{aligned} \tag{3}$$

Because the signal and the noise are mutually independent, the covariance matrix can be divided into the noise space and the signal space with EVD, which is shown in (4) and (5), where $\mathbf{\Lambda}$ means a diagonal matrix which is composed of the eigenvalue of $\mathbf{R_{uu}}$, and $\mathbf{q}_i$ is the eigenvector associated with the eigenvalue ς_i.

$$\mathbf{R_{uu}} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^H \tag{4}$$

where

$$\begin{aligned}\mathbf{\Lambda} &= diag\begin{bmatrix}\varsigma_0 & 0 & \cdots & 0 \\ 0 & \varsigma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varsigma_{m-1}\end{bmatrix} \\ \mathbf{B} &= \begin{bmatrix}\mathbf{q}_0 & \mathbf{q}_1 & \cdots & \mathbf{q}_{m-1}\end{bmatrix}\end{aligned} \tag{5}$$

using (4) and (5), we can observe

$$\begin{aligned}&|\mathbf{R_{uu}} - \varsigma_i\mathbf{I}| = 0 \\ &|\mathbf{A}\mathbf{R_{ss}}\mathbf{A}^H + \sigma_n^2\mathbf{I} - \varsigma_i\mathbf{I}| = 0 \\ &|\mathbf{A}\mathbf{R_{ss}}\mathbf{A}^H + (\sigma_n^2 - \varsigma_i)\mathbf{I}| = 0\end{aligned} \tag{6}$$

For only one source signal, the eigenvalue of the covariance matrix can be rewritten as

$$\mathbf{\Lambda} = diag\begin{bmatrix}\varsigma_s + \sigma_n^2 & 0 & \cdots & 0 \\ 0 & \sigma_n^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2\end{bmatrix} \tag{7}$$

where ς_s denotes the only non-zero positive eigenvalue of the covariance matrix. For i is more than or equal to one, we have

$$\begin{aligned}\mathbf{R_{uu}}\mathbf{q}_i &= \left[\mathbf{A}\mathbf{R_{ss}}\mathbf{A}^H + \sigma_n^2\mathbf{I}\right]\mathbf{q}_i \\ &= \varsigma_i\mathbf{q}_i \\ &= \sigma_n^2\mathbf{q}_i\end{aligned} \tag{8}$$

we can find that

$$\mathbf{A}\mathbf{R_{ss}}\mathbf{A}^H\mathbf{q}_i = 0 \tag{9}$$

According to (9), the eigenvector associated with the $m-1$ lowest eigenvalues of $\mathbf{R_{uu}}$ and the vector corresponding to the actual TDOA (see (10)) are orthogonal. Finally we can obtain the cost function and the angle of the source signal in (11).

$$\begin{aligned}&\mathbf{V}_n = \begin{bmatrix}\mathbf{q}_2 & \mathbf{q}_3 & \cdots & \mathbf{q}_{m-1}\end{bmatrix} \\ &\mathbf{A}^H(\theta)\mathbf{V}_n\mathbf{V}_n^H\mathbf{A}(\theta) = 0\end{aligned} \tag{10}$$

$$\begin{aligned}\theta_{MUSIC} &= \arg\max_{\theta} P_{COST}(\theta) \\ P_{COST}(\theta) &= \frac{1}{\mathbf{A}^H(\theta)\mathbf{V}_n\mathbf{V}_n^H\mathbf{A}(\theta)}\end{aligned} \tag{11}$$

3. Proposed system

The proposed speech SSL system is shown in Figure 1, which can be divided into two parts: The first one is the proposed pre-processing scheme, which contains the linear phase-difference approximation, covariance matrix reconstruction, and frequency bin selection; the second one is the DOA estimation. For the DOA evaluation, we utilize the MUSIC and the minimum variance distortionless response (MVDR) [7] algorithms to estimate the source direction in our proposed system.

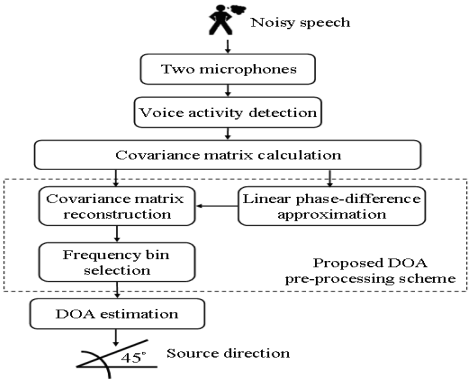

Figure 1: *Proposed far-field speech SSL system.*

3.1. Linear phase-difference approximation

3.1.1. Phase-difference line

We suppose that the sound signal are collected with two microphones in noiseless environment, and transform the two received signals into the frequency domain. The observed signals can be written as

$$\begin{aligned}\mathbf{U}(f) &= \begin{bmatrix}U_1(f) & U_2(f)\end{bmatrix}^T \\ &= \mathbf{A}(f)S(f) \\ &= \begin{bmatrix}1 & e^{-j2\pi f\tau}\end{bmatrix}^T S(f)\end{aligned} \tag{12}$$

where $U_n(f)$, $S(f)$, $\mathbf{A}$, f, and τ are, respectively, the observed signal, the source signal, steering vector, frequency, and delay time. The covariance matrix of the received signals can be defined in (13), where H denotes the Hermitian operation, and $E\{\bullet\}$ represents the expectation value of the matrix.

$$\begin{aligned}\mathbf{R_{uu}} &= E\{\mathbf{U}\mathbf{U}^H\} \\ &= E\left\{\begin{bmatrix}U_1U_1^H & U_1U_2^H \\ U_2U_1^H & U_2U_2^H\end{bmatrix}\right\}\end{aligned} \tag{13}$$

the components of $U_1U_2^H$ and $U_2U_1^H$ are shown as

$$\begin{aligned}U_1U_2^H &= S(f)^2e^{j2\pi f\tau} \\ U_2U_1^H &= S(f)^2e^{-j2\pi f\tau}\end{aligned} \tag{14}$$

According to (14), we can calculate the phase-difference for each frequency and represent these phase-differences as a linear line [6], which is described in Figure 2. Nevertheless, the observed phase-difference line does not have the linearity property in noisy environment (see details in Figure 3). In order to reduce the influences of ambient noise and the

reverberation, we propose an approximation procedure, which can make the observed phase line close to the theoretical one.

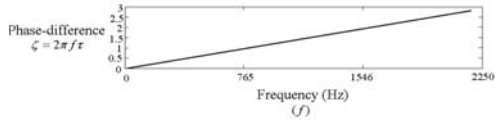

Figure 2: *Theoretical phase-difference line.*

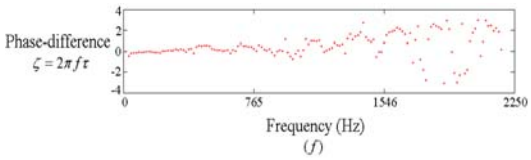

Figure 3: *Observed noisy phase-difference line.*

3.1.2. Linear phase approximation procedure

Figure 4 is the procedure of the proposed linear phase line approximation, where $\zeta_f^{original}$, $\zeta_f^{regression}$, and $\zeta_f^{updated}$ are, respectively, the original phase, the estimated phase after regression, and the final selected phase.

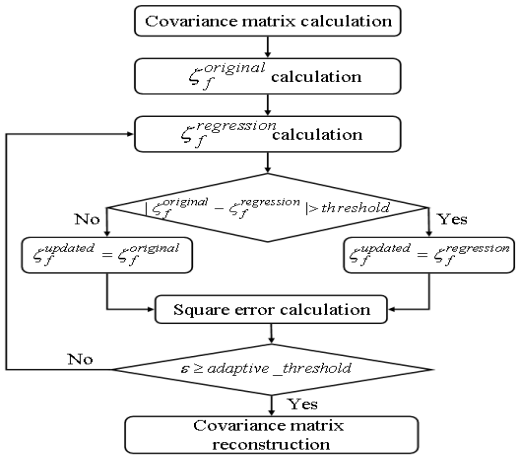

Figure 4: *Linear phase line approximation procedure.*

After the linear regression calculation, we can estimate the square error ε. Since the value of the square error is distinct from the different sound source direction, we utilize another threshold value *adaptive_threshold* to appropriately adjust the regression phase line. Figure 5 shows the phase-difference line after the proposed phase approximation.

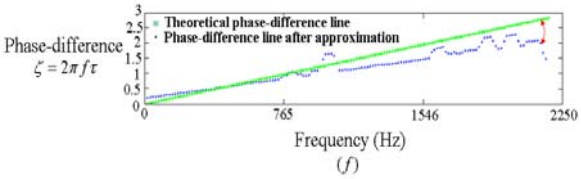

Figure 5: *Phase-difference line after approximation.*

3.2. Covariance matrix reconstruction

The objective of the covariance matrix reconstruction is to reconstruct the new covariance matrix with $\zeta_f^{updated}$, which is described in Section 3.1.2. The new covariance matrix can be given by (13), where the two components of $U_1U_2^H$ and $U_2U_1^H$ are shown as

$$U_1U_2^H = S(f)^2 e^{j\zeta_f^{updated}}$$
$$U_2U_1^H = S(f)^2 e^{-j\zeta_f^{updated}} \quad (15)$$

3.3. Frequency bin selection

Although speech is a wideband signal, not all of the frequency bins contain complete speech information in the frequency domain. Hence, we propose a frequency bin selection technique, which adopts the eigen signal-to-noise ratio (Eigen SNR), to discover which frequency bin includes the major speech signal property.

Figure 6 clarifies the procedure of proposed frequency bin selection, which principally consists of the EVD and the Eigen SNR calculation. When finishing the EVD of the covariance matrix (see (16)), we can obtain the eigenvalues of the source signal and the noise (i.e., ς_s and ς_n) from the covariance matrix $\mathbf{R_{uu}}$. The Eigen SNR is formulated in (17); the larger Eigen SNR value represents more speech information than smaller one.

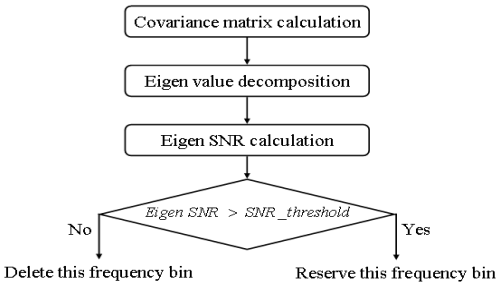

Figure 6: *Frequency bin selection procedure.*

$$\mathbf{R_{uu}} = \mathbf{B}\boldsymbol{\Lambda}\mathbf{B}^H$$
$$\boldsymbol{\Lambda} = diag\begin{bmatrix} \varsigma_s & 0 \\ 0 & \varsigma_n \end{bmatrix} \quad (16)$$
$$\mathbf{B} = \begin{bmatrix} \mathbf{q}_s & \mathbf{q}_n \end{bmatrix}$$

$$Eigen\ SNR = \frac{\varsigma_s}{\varsigma_n} \quad (17)$$

4. Experiments

4.1. Experimental setup

The detailed specification of the experimental environment is illustrated in Table 1, which contains the descriptions of microphone, test distance, test speaker, noise, recorded speech, and test direction.

Table 1. *Experimental setup specification.*

Specification item	Experimental setup specification
Microphone type	Omni-directional
Number of microphones	2
Microphone spacing	0.08 m
Sampling frequency	8 kHz
Distance (speaker & mic.)	3 m
Number of test speakers	11 (4 females, 7 males)
Noise type	White noise
SNR value of speech	10, 20, and 30 dB
Length of the test speech	Less than 3 seconds
Test direction of speech	60°, 75°, 90°, 105°, and 120°

4.2. Evaluation Results

In the experimental results, Figure 7 is the performance of MUSIC approach and our proposed system, which include the average degree errors with several test speeches (difference SNR values). About Figure 7, the horizontal axis is the sound direction, and the ordinate axis is the average error. It is clear to see that the proposed system is superior to the MUSIC approach; especially for 60 degrees and 120 degrees. The entire results are described in Table 2; the total average errors can be decreased by about 7.5 degrees. Figure 8 shows the performance of MVDR method and our proposed system. The total average errors can be reduced by about 5 degrees; the detailed results are described in Table 3.

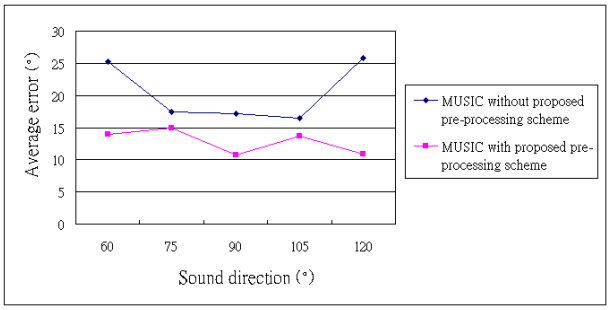

Figure 7: *Average errors of different directions.*

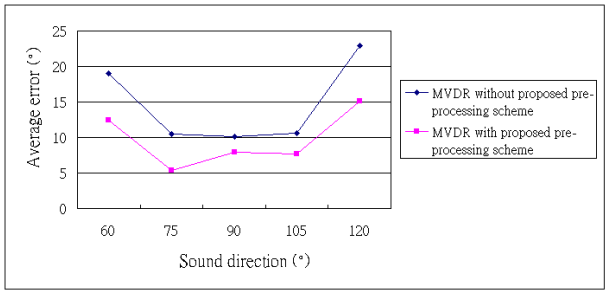

Figure 8: *Average errors of different directions.*

Table 2. *Comparison of average errors.*

Sound direction (°)	MUSIC without proposed pre-processing scheme	MUSIC with proposed pre-processing scheme
60	25.22°	13.96°
75	17.48°	14.89°
90	17.20°	10.80°
105	16.50°	13.61°
120	25.88°	10.94°
Average (°)	20.46°	12.84°

Table 3. *Comparison of average errors.*

Sound direction (°)	MVDR without proposed pre-processing scheme	MVDR with proposed pre-processing scheme
60	19.08°	12.44°
75	10.54°	5.36°
90	10.16°	7.89°
105	10.62°	7.63°
120	22.93°	15.12°
Average (°)	14.67°	9.69°

In the end, for the purpose of comparing the performances with several SNR values, Figure 9 represents the performances of MUSIC approach and the proposed system, which are included the average errors with the particular angle (90 degrees) and different SNR values. The detailed results are shown in Table 4, which also confirm that our proposed system is effective in reducing the noise influence.

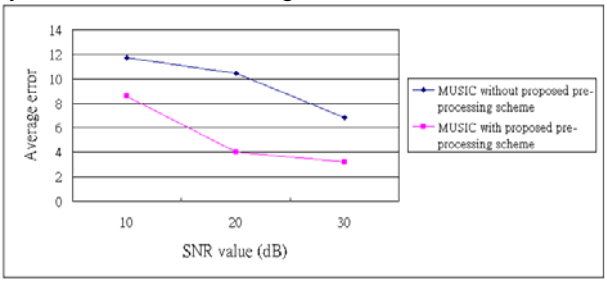

Figure 9: *Average errors of different SNRs.*

Table 4. *Average errors of different SNRs.*

SNR value (dB)	MUSIC without proposed pre-processing scheme	MUSIC with proposed pre-processing scheme
10	11.75°	8.60°
20	10.45°	4.05°
30	6.85°	3.20°

5. Conclusions

In this study, we proposed a far-field speech SSL system used in real acoustic environments. We utilize a proposed pre-processing scheme, which contains three stages: Linear phase-difference approximation, covariance matrix reconstruction, and frequency bin selection, to enhance localization ability. The experimental results show that our system can relieve the influences of the ambient noise and reverberation for several different sound directions and SNR values. Besides, this system can decrease the average degree errors about 5 to 7.5 degrees. It clearly demonstrates the performance of the proposed system and its feasibility.

6. References

[1] Ross, M., Shaffer, H., Cohen, A., Freudberg, R. and Manley, H., "Average magnitude difference function pitch extractor", IEEE Trans. Acoustics, Speech and Signal Proc., 22(5):353-362, 1974.

[2] Knapp, C. and Carter, G., "The generalized correlation method for estimation of time delay", IEEE Trans. Acoustics, Speech and Signal Proc., 24(4):320-327, 1976.

[3] Schmidt, R., "Multiple emitter location and signal parameter estimation", IEEE Trans. Antennas and Propagation, 34(3):276-280, 1986.

[4] Pillai, S.U. and Kwon, B.H., "Forward/backward spatial smoothing techniques for coherent signal identification", IEEE Trans. Acoustics, Speech and Signal Proc., 37(1):8-15, 1989.

[5] K. Cho, T. Nishiura and Y. Yamashita, "Robust speaker localization in a disturbance noise environment using a distributed microphone system", IEEE Symp. Chinese Spoken Language Proc., 209-213, 2010.

[6] Danfeng Li and Levinson, S.E., "A linear phase unwrapping method for binaural sound source localization on a robot", IEEE Conf. Robotics and Automation, 19-23, 2002.

[7] Capon, J., "High-resolution frequency-wavenumber spectrum analysis", Proc. IEEE, 57(8):1408-1418, 1969.