TWO-CHANNEL DOA ESTIMATION USIGN FREQUENCY SELECTIVE MUSIC ALGORITHM WITH A PHASE COMPENSATION IN REVERBERANT ROOM

Jae-Mo Yang[1], Min-Seok Choi[2] , Hong-Goo Kang[3]

Yonsei Univ. Dept. of Electrical & Electronic Eng., Seoul, 120-749, Republic of Korea
{Jaemo2879[1], zzugie[2], hgkang[3]}@dsp.yonsei.ac.kr

ABSTRACT

This paper proposes a robust two-channel frequency selective multiple signal classification (MUSIC) method to find a direction of arrival (DoA) information of speech signal. To overcome a phase distortion caused by reverberation and background noise in real acoustic room environments, we adopt a least square (LS)-based phase estimation method. In the phase compensation stage, distorted phases are replaced by estimated phases to enhance the accuracy of covariance matrix needed for the eigen-decomposition of the MUSIC method.

Simulation results verify that the proposed algorithm shows much higher estimation accuracy than conventional one while its complexity can be reduced by the frequency selection method.

Index Terms— multiple signal classification(MUSIC), direction of arrival(DoA) estimation.

1. INTRODUCTION

In a hands-free speech communication situation, a microphone array system has been widely used to acquire a high quality target speech. However, the beamforming or source separation algorithms with microphone arrays generally need to the information about the location of user or a direction of target signal arrival (DoA) [1]. The DoA estimation method is generally classified into three approaches, such as time delay estimation (TDE) [2], steered response power (SRP) [3], and eigen-decomposition based methods [4][5]. Typical methods commonly need a lot of microphones to correctly estimate DoA in real room conditions where reverberation often huddles for processing. Using multiple microphones is impractical in hand-held device environments, however, because of its physical size, cost and huge computational load. Several researchers have proposed methods using two microphones [6][7].

Hioka modified the multiple signal classification (MUSIC) for two-channel microphones by virtually constructing multichannel data with harmonics in vowel speech sound [7]. It worked as a frequency domain narrow-band multi-channel array data. Since the method requires harmonics, however, they only tested the system with five vowels (/a/ /e/ /i/ /o/ /u/) which have clear harmonic components. It also requires long interval of stationary regions in reverberation condition to calculate an accurate covariance matrix. However, the two assumptions, harmonics and long interval of stationarity, cannot be satisfied in voice communication situations practically, because we only can utilize a small portion of speech sound in the real-time system. If the small portion of speech segments is only allowed to use, we cannot avoid phase distortion caused by reverberation and background noise [8].

The main contribution of this paper is to overcome the phase distortion problem of the two-channel MUSIC algorithm using short time length of segments. By utilizing the phase variances of each frequency bins, we design a least square criterion to estimate phase terms. Specifically, we use phase variance as an weighting factor for phase estimation. The phase variances are calculated from the phase terms within a couple of neighborhood frames, which is utilized as weighting values at the phase estimation stage. The suspicious phases including distortions are replaced by estimated phases, which improves the accuracy of covariance matrix. Finally, DoA is calculated by applying the MUSIC method to the enhanced covariance matrix.

2. SIGNAL MODELING

In following discussion, we assume that signals received by two channel microphones, $x(n)$ and $y(n)$, are modeled as

$$x(n) = s(n) + n_x(n) + \sum_i \alpha_i s(n - \tau_i^{'}), \qquad (1)$$

$$y(n) = s(n-\tau) + n_y(n) + \sum_i \beta_i s(n - \tau_i^{''}),$$

where $s(n)$ is a target signal, $n_x(n)$ and $n_y(n)$ are independent white noise terms. The rightmost terms in both equations represent reverberation terms, where α_i and β_i are attenuation factors normally less than one. The time difference of two signals τ , relate to the arrival angle θ, is defined as $\tau(\theta) = d\sin(\theta)/c$, where d is a distance between two microphones and c is the sound velocity in the air. $\tau_i^{'}$ and $\tau_i^{''}$ are time differences caused by reverberation.

978-1-4244-2241-8/08/$25.00 ©2008 IEEE

Fourier transform of $x(n)$, $y(n)$ and their cross spectrum $G_{xy}(n)$ are represented by

$$X(\omega) = S(\omega) + N_x(\omega) + \sum_i \alpha_i S(\omega) e^{-j\omega\tau_i'}, \tag{2}$$

$$Y(\omega) = S(\omega) e^{-j\omega\tau} + N_y(\omega) + \sum_i \beta_i S(\omega) e^{-j\omega\tau_i''},$$

and

$$G_{xy}(\omega) = X^*(\omega) \times Y(\omega) = |S(\omega)|^2 e^{-j\omega\tau} + Z(\omega), \tag{3}$$

where

$$Z(\omega) = \sum_i \beta_i |S(\omega)|^2 e^{-j\omega\tau_i''} + \sum_i \alpha_i |S(\omega)|^2 e^{-j\omega(\tau - \tau_i')} \tag{4}$$

$$+ \sum_i \sum_j \alpha_i \beta_j |S(\omega)|^2 e^{-j\omega(\tau_j'' - \tau_i')}.$$

The term, $Z(\omega)$ is generated by correlation between target signal and its reverberation. If we ignore the reverberation effect of $Z(\omega)$ for simplicity, the cross spectrum becomes

$$G_{xy}(\omega) = |S(\omega)|^2 e^{-j\omega\tau}. \tag{5}$$

3. VIRTUAL MUSIC METHOD

To virtually generate a multi-channel data, we have to select several principle frequency bins that contain exact phase information. It is reasonable that frequency bins with higher power are more robust to phase distortion caused by reverberation and background noise [9]. An example given in Fig. 1 shows the feasibility of the approach. In the figure, vertical arrows depict phase distortion at the low power frequency bin. Therefore, it is clear that we had better select the candidate bins having higher power than others.

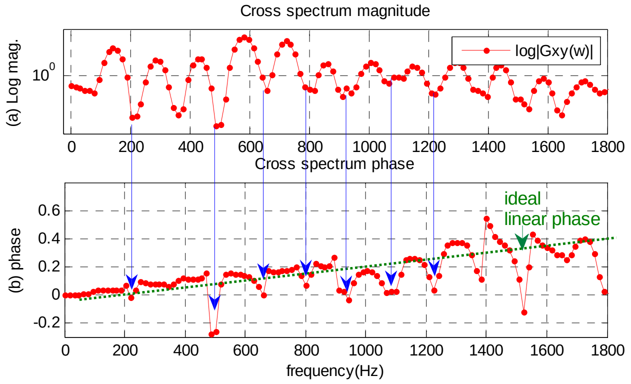

Fig. 1. Examples of cross spectrum and phase terms
(a) Cross spectrum (b) Phase(DoA=10°)

After selecting N frequency bins, the cross spectrum vector is reconstructed by

$$\mathbf{G}_{xy} = \left[G_{xy}(\omega_1) \quad \cdots \quad G_{xy}(\omega_N) \right]^T, \tag{6}$$

where subscript "T" denotes the matrix transpose. In the previous method proposed by Hioka [7], they used harmonics of fundamental frequency ω_0 to select principle frequency bins.

$$\mathbf{G}_{xy} = \left[G_{xy}(\omega_0) \quad G_{xy}(2\omega_0) \quad \cdots \quad G_{xy}(N\omega_0) \right]^T. \tag{7}$$

And then some harmonics whose powers are less than threshold are thrown away [7]. In the approach however, the system performance becomes too sensitive to the estimation accuracy of fundamental frequency because intervals between harmonic bins are not consistent. Since the performance of the frequency selection method is highly related to the accuracy of phase information, the method needs to be improved.

In this paper, we select candidate frequency bins by introducing a peak detection method. In other words, we select frequency bins which have comparatively higher power than neighborhood bins. The proposed approach is much simpler and useful at the following phase estimation stage because it increases the accuracy of phase information. Some frequency bins having phase distortion can be thrown away after performing a phase estimation process.

Using the fact that the cross spectrum vector can be considered as multiple number of virtually generated narrow band signals, the covariance matrix becomes

$$\mathbf{R}_{xy} = E[\mathbf{G}_{xy} \mathbf{G}_{xy}^H]. \tag{8}$$

Finally, DoA is estimated using the MUSIC method [4]

$$DOA_{MUSIC}(\theta) = \arg\max_{\theta} \frac{1}{e_s^H(\theta) \sum e_n e_n^H e_s(\theta)}, \tag{9}$$

where e_s and e_n are eigenvectors of signal and noise components, respectively and "H" denotes hermitian transpose.

4. PROPOSED METHOD

4.1. Phase estimation

Assuming that the selected N high power frequency bins have relatively correct phase information, more correct phase can be estimated if we apply an LS-based estimation method. Let the phase ϕ_i of i-th frequency bin be represented by

$$\phi_i = \omega_i \tau + 2\pi k_i, \quad i = 1, \ldots, N, \tag{10}$$

where τ is a linear tilt that represents time delay and $2\pi k_i$ is a unwrapping factor. After ignoring a unwrapping factor $2\pi k_i$ for simplicity, we construct a merit function with regard to τ. Expansion of whole equation to τ and $2\pi k_i$ is not much different from τ only case [10]. The merit function is represented by

$$\chi^2(\lambda) = \sum_{i=1}^{N} \left[\frac{\phi_i - \omega_i \tau_{LS}}{\sigma_i} \right]^2, \tag{11}$$

where σ_i is an uncertainty associated with each measurement ϕ_i and N is the total number of components within this segment. The least square solution of τ can be obtained by obtaining a minimum point of the equation (11) where, the derivative of $\chi^2(\lambda)$ with respect to τ equals to zero.

$$\frac{\partial \chi^2(\lambda)}{\partial \tau} = -2\sum_{i=1}^{N}\frac{\omega_i(\phi_i-\omega_i\tau_{LS})}{\sigma_i^2}=0. \qquad (12)$$

The solution of equation (12) becomes

$$\tau_{LS} = \left(\sum_{i=1}^{N}\frac{\omega_i\phi_i}{\sigma_i}\right)/\left(\sum_{i=1}^{N}\frac{\omega_i^2}{\sigma_i}\right). \qquad (13)$$

The uncertainty factor $1/\sigma_i$ is a weighting value of the estimator. In a conventional method, the power of frequency bin is commonly used as a weighting factor [9]. However, phase information is more important than power of each bin in this case. Assuming that phase terms do not changing much within a couple of neighborhood frames, we set the weighting values related to the phase variance of each frame shift. An example of this estimation result is depicted as triangular markers in Fig. 2.

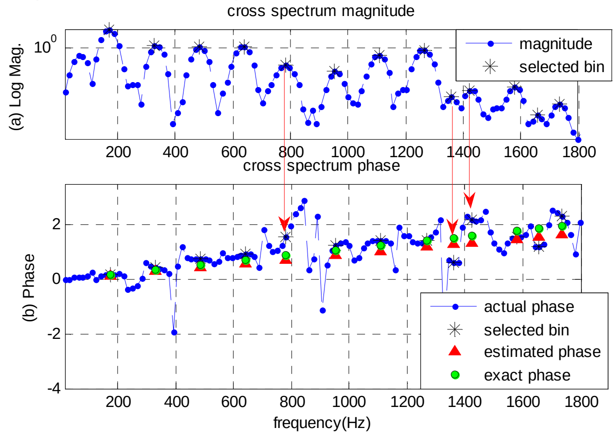

Fig. 2. Examples of phase distortion in three consecutive frames
(a) Cross spectrum magnitude (b) Cross spectrum phase

4.2. Phase compensation

Phase distortion is one of the most serious problems in DoA estimation. In conventional methods, smoothing is commonly used to reduce the phase distortion, however they need a lot of microphones for spatial smoothing or a lot of spectral resources for spectral smoothing [8]. We introduce a criterion of dropping distorted phase components, and replace the smoothing process by a phase compensation process within short time speech intervals. In Fig. 2 three actual phases pointed by vertical arrows are considered as distorted so that the distorted phases are replaced by estimated phases marked by (Δ).

Fig. 3 depicts a whole system block diagram. The main contribution of the proposed method is the LS phase estimation and phase compensation blocks.

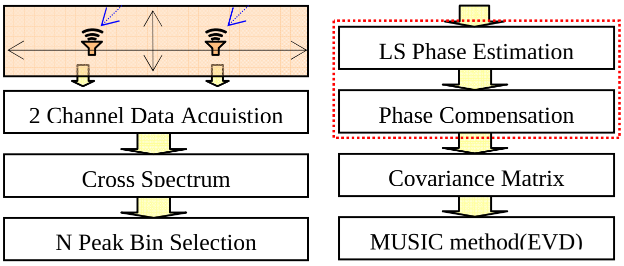

Fig. 3. An improved DoA estimation system

5. EXPERIMENTAL RESULTS

To verify the performance of the proposed method, we first perform a computer simulation with one word of speech sound having 3 second time interval. Simulation set up is summarized in Table 1.

Table 1. Summarizes detailed set up of the simulation

Sampling frequency	8kHz
Distance of microphones	0.08m
Window / length / overlap	Hamming / 256 point / 50%
FFT size	512 point

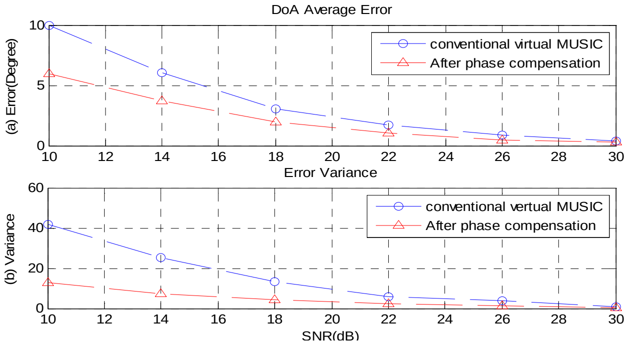

Fig. 4. Average error of DoA and its variance
(a) DoA average error (b) Error variance

The result given in Fig. 4 shows that the conventional virtual MUSIC method has poor performance in the low SNR conditions because the smoothing approach is not effective in this simulation condition. We can improve the estimation result by adopting the proposed phase compensation method even with insufficient resources conditions.

We have also evaluated the performance of the proposed algorithm in the real conference room ($8\times10\times3[m^3]$). Target signal consist of a 10 second time interval of real conversational speech including low frequency fan noise and acoustic reverberation. The distance between talker and microphone is 1.5m. System environment is same as the first simulation described in Table 1. We performed two experiments with a fixed talker at zero degree and a moving talker from 40° to -40° . The results are depicted in Fig. 5 and Fig. 6, respectively.

In Fig. 5, we can observe serious DoA errors with the conventional method, especially the ones represented by the arrow. After applying the proposed phase compensation, most serious estimation errors are removed.

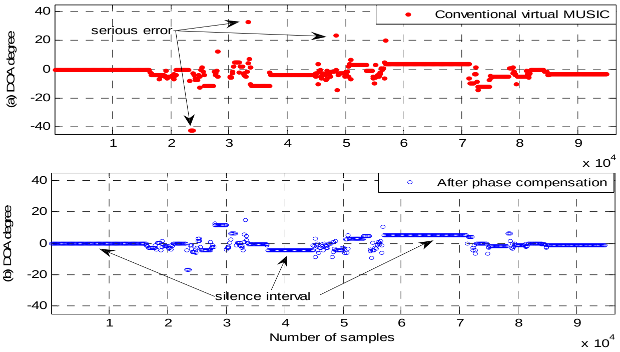

Fig. 5. Estimated DoA angle(fixed talker)
(a) Conventional virtual MUSIC (b) After phase compensation

Table. 2. Performance comparison

	Average angle	Deviation
Conventional	-4.17	4.94
Proposed	-1.04	2.95

Table 2 shows an average DoA result of the experiment. The improvement of our proposed method compared to conventional one is 40.3% at its deviation. The proposed algorithm can be also implemented into a real-time system because its complexity is pretty low.

In Fig. 6, we can also confirm that DoA error is reduced in the moving talker case and the final results become more closer to its average DoA angle by adopting the phase compensation method.

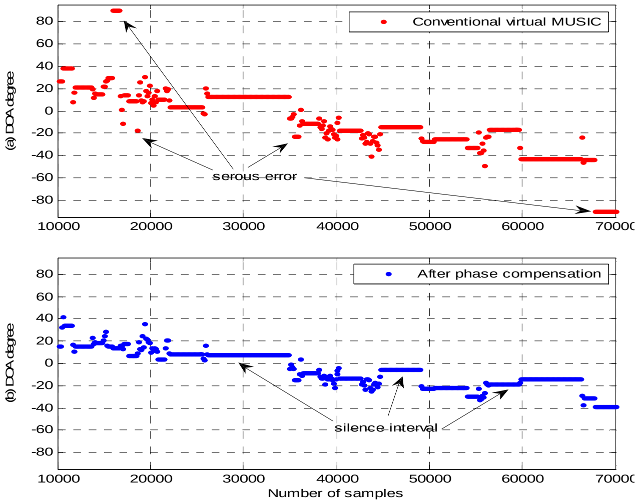

Fig. 6. Estimated DoA angle(moving talker)
(a) Fixed talker (b) Moving talker from 40° to -40°

6. CONCLUSTION

In this paper, we proposed a robust two-channel MUSIC method with a phase compensation technique. In conventional methods, large time intervals of multiple speech frames are needed to reduce phase distortion caused by the reverberation and background noise. However a cost problem always exists in this approaches, and the performance of the eigen-decomposition based methods becomes poor where the reverberation and background noise exist. The proposed phase compensation method correctly estimates DoA angles even with a small portion of speech segments in real room condition with only two microphones. In addition, we can avoid serious DoA errors caused by the phase distortion.

We verified the performance improvement of the proposed method both in computer simulation and in real room environments.

11. REFERENCES

[1] Lo. D. Goubran, R. A. Dansereau, R. M. Thompson, G. Schulz, “Robust joint audio-video localization in video conferencing using reliability information”, *Instrumentation and Measurement, IEEE Transactions*, Vol. 53, Issue 4, Page(s):1132 – 1139, Aug. 2004.

[2] C. H. Knapp, G. C. Carter, “The generalized correlation method for estimation of time delay”, *IEEE Trans. Acoust. Speech Signal Process., ASSP*, Vol.ASSP-24, NO. 4, 1976.

[3] Dmochowski J. P., Benesty J., Affes S., “A generalized steered response power method for computationally viable source localization”, *IEEE Trans. Audio, Speech, and Language Processing*, Vol PP, Issue99, pp.1-17, 2007.

[4] R.O. Schmidt., “A signal subspace approach to multiple emitterlocation and spectral estimation”, *Ph. D. dissertation*, Stanford Univ., Stanford, CA, 1981.

[5] R. H. Roy, "ESPRIT-Estimation of signal parameters via rotational invariance techniques." *Ph.D. dissertation, Stanford Univ.*, Stanford, CA, 1987.

[6] Y. Nagata, T. Fujioka, and M. Abe, “ Two dimentional DOA estimation of sound sources based on weighting wiener gain exploiting two directional microphones”, *IEEE Trans. Audio, Speech, and Language Processing*, Vol 15, NO. 2, pp.416-429, Feb. 2007.

[7] Y. Hioka, Y. Koizumi and N. Hamada, “Improvement of DOA estimation using virtually generated multichannel data from two-channel microphone array”, *Journal of Signal Processing*, Vol. 7, No. 1, pp.105–109, 2003.

[8] T. J. Shan, M. Wax, T. Kailath, “On spatial smoothing for estimation of coherent signals”, *IEEE Trans. On Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, pp. 802-811, Aug. 1985.

[9] M. S. Brandstein, “A framework for speech source localization using sensor arrays,” *Ph.D. thesis*, Dept. Elec., Eng., Brown University, Providence, RI, May 1995.

[10] Li, D. Levinson, S. E., “A linear phase unwrapping method for binaural sound source localization on a robot”, *IEEE International Conference on Robotics and Automation*, VOL 1, pages 19-23, 2002.