SOURCE AND DIRECTION OF ARRIVAL ESTIMATION BASED ON MAXIMUM LIKELIHOOD COMBINED WITH GMM AND EIGENANALYSIS

R. Nishimura[*]

National Institute of Information and Communications Technology
Resilient ICT Research Center
2-1-3 Katahira, Aoba-ku, Sendai, Japan

Y. Suzuki

Tohoku University
Research Institute of Electrical Communication
2-1-1 Katahira, Aoba-ku, Sendai, Japan

ABSTRACT

A method is proposed for estimating the source signal and its direction of arrival (DOA) in this paper. It is based on ML estimation of the transfer function between microphones combined with the EM algorithm for a Gaussian Mixture Model (GMM), assuming that the signal is captured at each microphone with delay corresponding to the traveling of sound and some decay. By this modeling, search for the maximum log-likelihood in the ML estimation can be realized simply by eigenvalue decomposition of a properly designed matrix. Computer simulation results show that the proposed method achieves SDR of greater than 10 dB regardless of amplitude difference between microphones and DOA estimation error of less than 8 degrees, on average. It is also shown that it can maintain high performance in various conditions.

Index Terms— ML estimation, Gaussian Mixture Model, Rayleigh quotient, sparseness, time-frequency masking

1. INTRODUCTION

A method of extracting the source signal of interest among other competing signals such as interfering speech and ambient noise is a crucially important objective of audio signal processing. Estimation of the direction of arrival of the source signal is also important for event detection on various scales from our daily lives to the infrasound radiated by natural events such as volcano explosion, earthquake, and tsunami. Therefore, for those tasks, various algorithms using multiple microphones have been proposed [1, 2]. It is often assumed that microphones are placed sufficiently close that the amplitude of the signal captured by each microphone is at the same level and that there is only an arrival time difference between them.

Specifically emphasizing speech sparseness under this assumption, Izumi *et al.* proposed a Blind Source Separation (BSS) method for a reverberant environment based on ML estimation combined with time-frequency masking [3]. An important shortcoming of the method is that it must use brute-force searching to execute ML estimation of the arrival time difference between microphones.

Maruyama *et al.* further developed the method and proposed an analytical update rule for Time Difference of Arrival (TDOA) estimation [4]. In this method, by modeling the delay as a function not only of the source direction but also of the frequency, an analytical update rule for the delay parameter is derived theoretically by differentiating the auxiliary function. The obtained function has a similar shape to that of a von Mieses distribution [5]. The method achieved drastic reduction in computational time by avoiding brute-force searching but still including an iteration process to execute the EM algorithm and assuming the fixed level difference between microphones.

Inspired by an earlier work [4], the present study takes another approach by which it is presumed that the signal is captured at each microphone not only with delay, but also with decay. This presumption is validated because applications we are considering are not only those for speech but also infrasound monitoring, where the signal has a very low frequency so that the microphones need to be placed distant for DOA estimation and the signal is expected to decay while traveling [6]. By this new modeling, the maximization of the log-likelihood can be realized directly by eigenvalue decomposition rather than by application of an EM algorithm.

2. CONVENTIONAL METHOD

We assume that the signal is sparse in the time-frequency domain. The time-frequency representation of the signal captured at the microphones is

$$\boldsymbol{x}_{f,t} = \boldsymbol{b}_{f,k} S_{f,t,k} + \boldsymbol{N}_{f,t}, \tag{1}$$

where $S_{f,t,k}$ represents the time-frequency representation of the source signal, k is the index of the source, and f and t respectively denote indices of the frequency and the time. If there is only phase difference and no level difference between

[*]This research and development work was supported by the MIC/SCOPE #171502001.

the microphones, then the transfer function $\boldsymbol{b}_{f,k}$ can be modeled as

$$\boldsymbol{b}_{f,k} = \begin{bmatrix} 1 & \exp(j2\pi f\delta_k) \end{bmatrix}^T, \tag{2}$$

where δ_k is the arrival time difference for source signal k. In addition, an additive noise $\boldsymbol{N}_{f,t}$ is assumed to be present: its covariance matrix is represented as $\boldsymbol{V}_f$. Under the assumption of the diffused noise, it is represented by

$$\boldsymbol{V}_f = \sigma_f^2 \begin{bmatrix} 1 & \mathrm{sinc}(2\pi f\kappa) \\ \mathrm{sinc}(2\pi f\kappa) & 1 \end{bmatrix}, \tag{3}$$

where κ is the ratio of the distance between microphones to the speed of sound. Let $\theta = \{\sigma_f^2,\ \delta_k,\ S_{f,t,k}\}$ be a parameter set to estimate. Then, the conditional probability of the observed signal becomes [7]

$$p(\boldsymbol{x}_{f,t}|k,\theta) = \frac{1}{\pi\sigma_f^2|\boldsymbol{V}_f|}\exp\left(-\frac{1}{\sigma_f^2}\boldsymbol{N}_{f,t}^H\boldsymbol{V}_f^{-1}\boldsymbol{N}_{f,t}\right). \tag{4}$$

Moreover, the auxiliary function of the problem is

$$\begin{aligned} Q(\theta|\theta') &= \mathrm{E}[\log p(\boldsymbol{x}_{f,t};\theta)|\theta'] \\ &= \sum_f\sum_t\sum_k m_{f,t,k}\log p(\boldsymbol{x}_{f,t}|k,\theta)p(k|\theta), \end{aligned} \tag{5}$$

where θ' is the parameter set in the previous iteration and $m_{f,t,k}$ is time-frequency mask updated by

$$m_{f,t,k} = p(k|\boldsymbol{b}_{f,t,k},\theta') = \frac{p(k|\theta')p(\boldsymbol{x}_{f,t}|k,\theta')}{\sum_k p(k|\theta')p(\boldsymbol{x}_{f,t}|k,\theta')}. \tag{6}$$

The maximum likelihood estimate of the parameter is obtained as

$$\delta_k = \arg\max_{\delta_k} Q(\theta|\theta'). \tag{7}$$

A shortcoming of the method in [3] is that brute-force searching is required to seek the optimum δ_k in calculating (7).

Assuming that the delay parameter δ_k is a function not only of source but also of frequency, Maruyama *et al.* derived a new update rule. Under this assumption, the likelihood function is represented by

$$\begin{aligned} p(\boldsymbol{x}_{f,t}|k,\theta) &= \frac{1}{\pi\sigma_f^2|\boldsymbol{V}_f|}\cdot\exp(C) \\ &\cdot\exp\left\{\frac{2|\xi_{f,t,k}||S_{f,t,k}|}{\sigma_f^2(1-\phi_f^2)}\cos(\psi_{S_k}-\psi_{\xi_k}-2\pi f\delta_{f,k})\right\}, \end{aligned} \tag{8}$$

where $\phi_f = \mathrm{sinc}(2\pi f\kappa)$, $\xi_{f,t,k} = \boldsymbol{x}_{R,f,t} - \phi_f(\boldsymbol{x}_{L,f,t} - S_{f,t})$, ψ_{ξ_k} and ψ_{S_k} respectively denote phases of ξ_k and S_k. Also, C is independent from $\delta_{f,k}$. By substituting (8) into (5) and by setting $\frac{\partial Q}{\partial\delta_{f,k}} = 0$, the update rule is obtained as

$$2\pi f\delta_{f,k} = \arctan\frac{\sum_t m_{f,t,k}|\xi_{f,t,k}||S_{f,t,k}|\sin(\psi_{S_k}-\psi_{\xi_k})}{\sum_t m_{f,t,k}|\xi_{f,t,k}||S_{f,t,k}|\cos(\psi_{S_k}-\psi_{\xi_k})}. \tag{9}$$

In addition to solving the ambiguity in phase, the following correction is applied:

$$\begin{cases} 2\pi f\delta_{f,k} \leftarrow 2\pi f\delta_{f,k} + \pi & \text{for} \quad \delta_{f,k} < 0,\ \frac{\partial^2 Q}{\partial\delta_{f,k}^2} \geq 0 \\ 2\pi f\delta_{f,k} \leftarrow 2\pi f\delta_{f,k} - \pi & \text{for} \quad \delta_{f,k} > 0,\ \frac{\partial^2 Q}{\partial\delta_{f,k}^2} \geq 0 \end{cases}. \tag{10}$$

3. PROPOSED METHOD

3.1. ML estimation of transfer function

We model the time-frequency representation of the observed signal as

$$\boldsymbol{x}_{f,t} = \boldsymbol{b}_{f,k}S_{f,t,k} + \boldsymbol{N}_{f,t}, \tag{11}$$

where

$$\boldsymbol{b}_{f,k} = \begin{bmatrix} a_{L,f,k}\exp(j2\pi f\delta_{L,f,k}) & a_{R,f,k}\exp(j2\pi f\delta_{R,f,k}) \end{bmatrix}^T. \tag{12}$$

This representation of the observed signal (11) is the same as (1), but the transfer function $\boldsymbol{b}_{f,k}$ of (2) is replaced by (12) to include signal level difference between microphones. Parameter sets $\{a_{L,f,k},\ a_{R,f,k}\}$ and $\{\delta_{L,f,k},\ \delta_{R,f,k}\}$ respectively represent the signal decay and delay. This new model allows the transfer function to be any arbitrary vector. Under the assumption that $\boldsymbol{N}_{f,t}$ is a complex Gaussian distribution of mean zero and covariance matrix $\boldsymbol{V}$, the log likelihood can be represented as

$$\begin{aligned} &\log p(\boldsymbol{x}_{f,t}|\delta_{\cdot,f,k}) \\ &= C - (\boldsymbol{x}_{f,t} - \boldsymbol{b}_{f,k}S_{f,t,k})^H\boldsymbol{V}_f^{-1}(\boldsymbol{x}_{f,t} - \boldsymbol{b}_{f,k}S_{f,t,k}) \\ &= C - \boldsymbol{x}_{f,t}^H\boldsymbol{V}_f^{-1}\boldsymbol{x}_{f,t} + \frac{|\boldsymbol{b}_{f,k}^H\boldsymbol{V}_f^{-1}\boldsymbol{x}_{f,t}|^2}{\boldsymbol{b}_{f,k}^H\boldsymbol{V}_f^{-1}\boldsymbol{b}_{f,k}}, \end{aligned} \tag{13}$$

where

$$C \equiv -\log(\pi) - \log|\boldsymbol{V}_f|. \tag{14}$$

In this derivation,

$$\frac{\partial\log p(\boldsymbol{x}_{f,t}|\delta_{\cdot,f,k})}{\partial S_{f,t,k}} = 0 \ \Leftrightarrow\ S_{f,t,k} = \frac{\boldsymbol{b}_{f,k}^H\boldsymbol{V}_f^{-1}\boldsymbol{x}_{f,t}}{\boldsymbol{b}_{f,k}^H\boldsymbol{V}_f^{-1}\boldsymbol{b}_{f,k}} \tag{15}$$

is used for the unknown parameter $S_{f,t,k}$. Maximization of the log-likelihood can be deduced to maximization of the third term of (13) because it is the only term that is dependent on $\delta_{\cdot,f,k}$.

Because a covariance matrix $\boldsymbol{V}_f$ is a Hermitian, so is its inverse $\boldsymbol{V}_f^{-1}$. Therefore, it can be decomposed by the spectral theorem for Hermitian matrices [8] as

$$\boldsymbol{V}_f^{-1} = \boldsymbol{U}_f\Lambda_f\boldsymbol{U}_f^H = \boldsymbol{W}_f^H\boldsymbol{W}_f, \tag{16}$$

where

$$\boldsymbol{W}_f \equiv \sqrt{\Lambda_f}\boldsymbol{U}_f^H \tag{17}$$

and Λ_f is a diagonal matrix with diagonals consisting of eigenvalues of the matrix $\boldsymbol{V}_f^{-1}$, whereas columns of matrix $\boldsymbol{U}_f$ are eigenvectors for the corresponding eigenvalues. Moreover, defined $\boldsymbol{c}_{f,k} = \boldsymbol{W}_f \boldsymbol{b}_{f,k}$, the third term of (13) can be rewritten as

$$\begin{aligned}\frac{|\boldsymbol{b}_{f,k}^H \boldsymbol{V}_f^{-1} \boldsymbol{x}_{f,t}|^2}{\boldsymbol{b}_{f,k}^H \boldsymbol{V}_f^{-1} \boldsymbol{b}_{f,k}} &= \frac{|\boldsymbol{b}_{f,k}^H \boldsymbol{W}_f^H \boldsymbol{W}_f \boldsymbol{x}_{f,t}|^2}{\boldsymbol{b}_{f,k}^H \boldsymbol{W}_f^H \boldsymbol{W}_f \boldsymbol{b}_{f,k}} \\ &= \frac{(\boldsymbol{b}_{f,k}^H \boldsymbol{W}_f^H \boldsymbol{W}_f \boldsymbol{x}_{f,t})(\boldsymbol{x}_{f,t}^H \boldsymbol{W}_f^H \boldsymbol{W}_f \boldsymbol{b}_{f,k})}{\boldsymbol{b}_{f,k}^H \boldsymbol{W}_f^H \boldsymbol{W}_f \boldsymbol{b}_{f,k}} \\ &= \frac{\boldsymbol{c}_{f,k}^H \boldsymbol{D}_{f,t} \boldsymbol{c}_{f,k}}{\boldsymbol{c}_{f,k}^H \boldsymbol{c}_{f,k}}, \end{aligned} \tag{18}$$

where

$$\boldsymbol{D}_{f,t} \equiv \boldsymbol{W}_f \boldsymbol{x}_{f,t} \boldsymbol{x}_{f,t}^H \boldsymbol{W}_f^H = \boldsymbol{y}_{f,t} \boldsymbol{y}_{f,t}^H, \tag{19}$$

$$\boldsymbol{y}_{f,t} \equiv \boldsymbol{W}_f \boldsymbol{x}_{f,t} = \sqrt{\Lambda} \boldsymbol{U}_f^H \boldsymbol{x}_{f,t}. \tag{20}$$

The right-hand side of (18) is a form of Rayleigh quotient. Therefore, its maximum value is obtained as the largest eigenvalue of the matrix $\boldsymbol{D}_{f,t}$ [9]. Letting $\tilde{\boldsymbol{c}}_{f,t}$ be the corresponding eigenvector for the maximum eigenvalue of $\boldsymbol{D}_{f,t}$, then the ML estimate of the transfer function $\boldsymbol{b}_{f,k}$ can be obtained, using (16), as

$$\boldsymbol{b}_{f,k} = (\boldsymbol{W}_f^H \boldsymbol{W}_f)^{-1} \boldsymbol{W}_f^H \tilde{\boldsymbol{c}}_{f,t} = \boldsymbol{V}_f \boldsymbol{W}_f^H \tilde{\boldsymbol{c}}_{f,t}. \tag{21}$$

It is further normalized by the component of one microphone to obtain the arrival time difference between the microphones as

$$\boldsymbol{b}_{f,k} \leftarrow \begin{bmatrix} 1 & (a_{R,f,k}/a_{L,f,k}) \exp\{j2\pi f(\delta_{R,f,k} - \delta_{L,f,k})\} \end{bmatrix}. \tag{22}$$

The transfer functions obtained by (22) are a mixture of all sources for the specified time frame. Therefore, it should be further segregated into each source component using, for example, a clustering algorithm.

3.2. Clustering of signal components

Under the assumption of speech sparseness, each time-frequency component belongs to only one source signal. Define the arrival time difference $\varphi_{f,k}$ as

$$\angle \boldsymbol{b}_{f,k} = 2\pi f(\delta_{R,f,k} - \delta_{L,f,k}) \equiv 2\pi f \varphi_{f,k} \tag{23}$$

and that of time frame t as $\varphi_{f,t,k}$. A subset $\{\varphi_{f,t,k} | f_L \leq f \leq f_H\}$ is used in the clustering based on the EM algorithm for the Gaussian mixture model (GMM). It is also possible to execute the GMM estimation frame-by-frame so that it can cope with time-varying DOAs, but it will worsen the EM algorithm convergence. Frequency is bounded between $\{f_L,\ f_H\}$ for good stability because, for low frequency, a slight phase difference might cause a large deviation in the arrival time difference because of the division by frequency. For high frequencies, because an ambiguity in phase might occur, f_H needs to be set to $1/(2\Delta t)$, where Δt is the maximum traveling time between the microphones. The number of Gaussian distributions should be set to the assumed number of sources plus one. This plus-one source is for ambient noise, which is expected to be omnidirectional and to have a distribution with a mean of zero. By performing EM algorithm for GMM, a set of estimates for mean and standard deviation $\{\mu_k, \sigma_k\}$ is obtained. The obtained means of the estimated Gaussian distributions are then sorted to resolve the permutation problem. Those for standard deviation are also permuted accordingly.

Using the estimated Gaussian distribution as probability, the time-frequency mask to $\boldsymbol{x}_{f,t}$ for each source signal is obtained as

$$m_{f,t,k} = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{ \frac{(\varphi_{f,t,k} - \mu_k)^2}{2\sigma_k^2} \right\} / \max_{f,t,k}[m_{f,t,k}]. \tag{24}$$

Consequently, the source signal is reconstructed as

$$S_{f,t,k} = m_{f,t,k} \frac{\boldsymbol{b}_{f,k}^H \boldsymbol{V}_f^{-1} \boldsymbol{x}_{f,t}}{\boldsymbol{b}_{f,k}^H \boldsymbol{V}_f^{-1} \boldsymbol{b}_{f,k}}. \tag{25}$$

This form is the same as in [4] but using (22) and (24) instead of (2) and (6), respectively, for $\boldsymbol{b}_{f,k}$ and $m_{f,t,k}$.

4. PERFORMANCE EVALUATION

4.1. Conditions

Computer simulations were conducted under the conditions described below. Two microphones were placed with distance of 10 cm. The directions of speakers were set to a 30 and 60 deg incident angle and 2 m distant from the center of the microphones, as portrayed in Fig. 1. The signal is captured by the microphones at a sampling rate of 16 kHz, with quantization of 16 bits. Amplitude of the signal for one microphone is gained by $(1 - \beta)$ where $\beta = 0.0,\ 0.1,\ 0.2,\ 0.3,\ 0.4$. Japanese speech uttered by five male and five female speakers was used. Spoken sentences differed among speakers. The averaged duration of the speech segments was 5.2 s. A white Gaussian noise signal was added to the speech signal to form the captured signal. Its level is adjusted appropriately to have a 10 dB or 20 dB signal-to-noise ratio (SNR) at the fixed-amplitude microphone rather than the one gained by $(1 - \beta)$. Finally, after the captured signal is segmented into frames using a Hanning window of 1024 points with 50% overlap, it is then transformed into the frequency domain using Discrete Fourier Transform (DFT) to obtain its time-frequency representation.

Computer simulations were conducted only for a single source case because our primary interest is in estimating both DOA and the source signals quickly. Although Blind Source

Separation (BSS) is the main purpose of [4], we used it as the conventional method because it has abilities of source signal estimation, DOA, and noise reduction, irrespective of the numbers of sources.

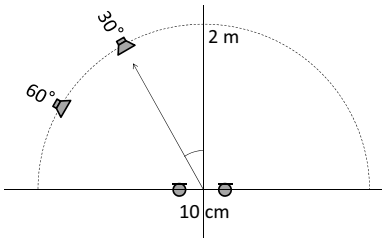
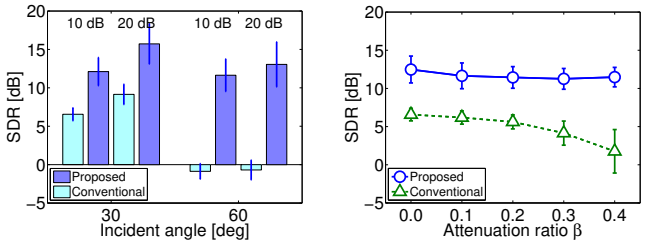


Fig. 1. Configuration of speaker and microphone positions.

4.2. Results

Performance in estimation of the sound source is evaluated in terms of the Signal to Distortion Ratio (SDR) [10]. Fig. 2 presents the results. In Fig. 2(a), circles are for the input SNR of 10 dB. Upside-down triangles are for 20 dB. The incident angles of 30 dB and 60 dB are represented respectively by solid and dashed lines. Error bars represent standard deviation. It is common for both methods that higher SDRs are obtainable for higher input SNRs. Results show that SDRs of the conventional method for the incident angle of 30 deg were approximately 6–9 dB. This result is comparable to experiment results described by [4]. As shown in Fig. 2(b), the SDR of the conventional method decreases as β increases, whereas that of the proposed method remains constant irrespective of the attenuation ratio. This is one benefit of the proposed method, which introduces additional parameters for the signal level.

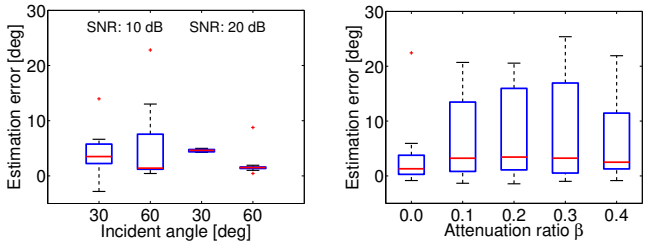


(a) Source position and SNR (b) Level difference

Fig. 2. SDR for different conditions in (a) source position and input SNR and the (b) signal level difference between microphones.

The drastic decrease in SDR for the larger incident angle on the conventional method shown in Fig. 2(a) was attributable the traveling time between two microphones, which exceeded the sampling rate.

Fig. 3 presents simulation results related to incident angle estimation. Other conditions are the same as Fig. 2. Only those of the proposed method are shown because the conventional method models the arrival time difference, which is used to calculate DOA, as a function not only of source direction but also frequency. It is therefore difficult to ascertain a specific DOA directly for each source. It is apparent from Fig. 3(a) that whiskers become shorter as the input SNR increases and that errors become smaller as the incident angle increases. Regarding other parameters, neither Fig. 3(a) nor Fig. 3(b) shows any remarkable tendency. Therefore, results suggest that the proposed method can achieve robust DOA estimation in various conditions.



(a) Source position and SNR (b) Level difference

Fig. 3. Estimation error of incident angle for different conditions in (a) source position and input SNR, and (b) the signal level difference between microphones.

The outliers suggest that a large estimation error might have occurred. The EM algorithm for GMM estimation sometimes did not converge within the specified maximum number of iterations, which might be the reason underlying the large estimation error. In addition, some bias exists in the resultant estimates of source DOA. A possible reason is that some isotropic error which occurred in the phase estimation produced non-isotropic error in the arrival time difference because of the division by frequency.

5. CONCLUSION

Assuming that both time and level differences exist among signals captured by multiple microphones, a new method is proposed for estimating both DOA and the source signal. The method is based on the ML estimation of time difference for each time-frequency component and EM algorithm for GMM. Computer simulations revealed that the method can function under various conditions. However, it is somewhat heuristic to regard a Gaussian distribution obtained by EM algorithm directly as a probability of the source signal. Development of a theoretical rule to ascertain the optimal time-frequency mask under the assumption considered here is left as a topic for future research.

6. REFERENCES

[1] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, Jul 1989.

[2] Y. Zhang and B. P. Ng, "MUSIC-like DOA estimation without estimating the number of sources," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1668–1676, March 2010.

[3] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2007, pp. 147–150.

[4] T. Maruyama, S. Araki, T. Nakatani, S. Miyabe, T. Yamada, S. Makino, and A. Nakamura, "New analytical update rule for TDOA inference for underdetermined bss in noisy environments," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 269–272.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[6] A. L. Pichon, E. Blanc, and A. Hauchecorne, Eds., *Infrasound Monitoring for Atmospheric Studies*, Springer, 2009.

[7] N. R. Goodman, "Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction)," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 152–177, 1963.

[8] G. Strang, *Linear Algebra and Its Applications*, Academic Press, 1976.

[9] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall International, Inc., second edition, 1991.

[10] T. Nishiura, S. Nakanura, and K. Shikano, "Speech enhancement by multiple beamforming with reflection signal equalization," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2001, pp. 189–192.