IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 30, 2021, accepted August 21, 2021, date of publication August 24, 2021, date of current version September 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3107412

Prefiltering Approach of Adaptive Eigenvalue Decomposition Method for an Improved Time Delay Estimation in Room Environment

YUN-HO SHIN[1] **AND JEUNG-HOON LEE**[2]

[1]Department of Safety Engineering, Chungbuk National University, Seowon-gu, Cheongju 28644, South Korea
[2]School of Mechanical Engineering, Changwon National University, Uichang-gu, Changwon 51140, South Korea

Corresponding author: Jeung-Hoon Lee (jhoonlee@changwon.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) under Grant 2021R1A2C1005962, and in part by the Korea Institute of Marine Science and Technology Promotion (KIMST) Grant 20210500.

ABSTRACT Time delay estimation is an essential step in sound source localization and beamforming systems, and an extensive amount of research has been performed on this subject. This task entails the accurate estimation of the relative time delay between two microphone signals originating from the same source. To overcome the limitations of the existing methods, the adaptive eigenvalue decomposition (AED) algorithm was developed for time delay estimation in reverberant acoustic environments in the early 2000s. This paper attempts to improve delay estimation performance using autoregressive model-based prefiltering for the AED algorithm. The proposed method establishes an autoregressive model of the room impulse response beforehand. Then, the model is utilized as a linear prediction filter to remove the reverberation component from the microphone signals, which improves the estimation capability of the algorithm. Monte Carlo experiments are performed to demonstrate the improved performance for various reverberation levels and signal-to-noise ratios. Provided that the noise level is moderate, the proposed method is shown to greatly increase the accuracy of the conventional approach in a highly reverberant room.

INDEX TERMS Time delay estimation, adaptive eigenvalue decomposition, autoregressive prefilter, source identification, model-based prefiltering technique.

I. INTRODUCTION

Time delay estimation (TDE), as an essential step for sound source localization and beamforming systems, has attracted an extensive amount of research interest [1]–[3]. This task entails the accurate estimation of the relative time delay between two microphone signals originating from the same source. It has many applications in various fields, including radar, sonar, seismology, geophysics, ultrasonics, communications, and more. Other recent applications involve performing tasks in room environments, such as localizing and tracking the active speaker in a teleconference system.

One of the most widely used techniques for TDE is generalized cross correlation (GCC), proposed in a landmark paper by Knapp and Carter [4]. The delay is given by the time lag at which the cross-correlation function between two received signals is maximized. This method fairly works well in the presence of moderate noise. However, its performance suffers when reverberation occurs in the room; it is prone to fail even in the presence of a weak echo. Although there exist a variety of GCC family methods, they cannot handle reverberation appropriately. This limitation can largely be attributed to the fact that most of these methods assume an ideal propagation model without reverberation, i.e., only a direct path between the signal source and the microphones. Recently, a study was conducted on an improved algorithm based on GCC using phase transform weights [5], [6], but the method was observed to be sensitive to noise. In addition, methods [7], [8] based on various information including microphone signals have also been studied in order to estimate the correct direction of a sound source under the assumption that the microphone array can be used, and methodologies [9], [10] for estimating various sources expected in real situations are the subject of ongoing research.

A new approach, referred to as adaptive eigenvalue decomposition (AED), was proposed by Jacob [11]. This method uses a more realistic signal model to consider reverberation in the room. Unlike in the GCC-based methods, impulse responses from the source to two microphones are identified using an adaptive algorithm that iteratively calculates

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

the eigenvector corresponding to the smallest eigenvalue of the covariance matrix. Then, the time delay is estimated as the time difference between dominant peaks of the two identified impulse responses, or as the time lag that maximizes the cross-correlation function between the two impulse responses. It is emphasized that the focus of the AED method lies in the estimation of the relative time difference between the main peaks (direct path) of the acoustic impulse responses, not in the exact estimation of the acoustic impulse response. As follow-up research, many studies have also been conducted to improve the robustness of TDE, adopting methods to reflect spatial diversity using two or more microphones [12]–[14], improving performance using GCC weightings [15], and using blind channel estimation [11], as well as considering reverberation and correlated noise fields [16]. Very recently, a performance improvement study using a sub-band approach [17] to improve TDE for speech sources using a multichannel sparse linear prediction algorithm [18] has been conducted. In addition to studies on source localization using AED and studies seeking to improve the robustness of these methods, research is continuously being conducted on related topics in various fields, including tracking the desired sound from microphone signals while removing noise from the measured signals [19], improving the computational speed of the adaptive algorithm to perform signal processing more quickly [20], [21], and performing TDE for watermarked audio signals [22].

Although the AED algorithm demonstrates the best performance among the available TDE techniques, including the GCC-based method [23], its validity holds only for mild reverberation conditions, and breaks down if the reverberation becomes severe. Thus, the current investigation attempts to improve the TDE performance of AED by adopting the relatively simple prefiltering approach, using known sound field information, unlike previous studies. More specifically, the reverberation structure is first estimated by performing autoregressive modeling of the room impulse response. Using the model as a linear prediction filter eliminates the reverberation from the microphone signals, leaving the direct component in the residual. Working with the residual rather than the original signal assists the AED in determining the direct path signals. Using AR modeling to determine the impulse response functions constitutes a task that is completely different from that which AED performs. As detailed in the next section, the AED algorithm was designed to highlight direct paths only, whereas the AR model describes the entire reverberation structure as faithfully as possible using the received signal.

The notion of prefiltering is not new; it is frequently exploited in the fault diagnosis of rotating machinery such as bearings and gears [24]–[27]. In such applications, an AR model is employed to predict the deterministic sinusoidal data or sharp spectral peaks corresponding to the rotational components and their harmonics. A fault signal (usually a series of impulses) is then extracted by removing the deterministic part, and can be further enhanced with subsequent post-processing. Likewise, the signal in a reverberant field can be regarded as being corrupted by the room resonance. Thus, we anticipate that the resonant components in in the microphone signal can be described well through the AR modeling of the room impulse response. That is, prefiltering using the AR model can also be applicable for the removal of room reverberation. In fact, some studies [28], [29] have claimed that dereverberation for the enhancement of speech signals can be achieved efficiently using the AR model rather than the moving-average (MA) model. Thus, the present work offers the distinct contribution of adopting the AR prefilter for the performance enhancement of AED, an approach that has not been attempted in previous studies.

The remainder of the paper is organized as follows. The next section first introduces the AED algorithm, as previously proposed by Jacob [11], then suggests an AR prefiltering method to improve the performance of the AED algorithm. Additional discussion on the parameter selection for the AED is provided as a complement of this work in Section II.A. The AR prefiltering method is further developed in Section II.B together with the selection of an appropriate order to optimize performance. Numerical experiments reported in Section III elucidate the benefits of the proposed method considering various room environment conditions and arrangements of microphones and sources. Finally, Section IV concludes this paper.

II. PROPOSED METHOD

A. OVERVIEW OF ADAPTIVE EIGENVALUE DECOMPOSITION METHOD

A simple propagation model for the TDE problem is given by

$$x_i(n) = \alpha_i s(n - \tau_i) + w_i(n), \tag{1}$$

where $x_i(n)$ (for $i = 1, 2$), is the signal output of the i-th microphone at time index n (= 0, 1, . . . , N-1), α_i is the attenuation factor along the propagation path, $s(n)$ is the source signal, τ_i is the time it takes to reach the i-th microphone from the source, and $w_i(n)$ is the zero-mean, uncorrelated, and stationary Gaussian noise component added to the i-th microphone. It is assumed that the spectral component of $s(n)$ is fairly broad-banded. Our aim is to estimate accurately the relative delay between the two microphone signals, i.e., $\tau = \tau_2 - \tau_1$, using the finite sets of observation samples of $x_1(n)$ and $x_2(n)$.

Equation (1) deals with an ideal situation in which the signal propagation from the source to each microphone takes place along a single direct path, and is no longer valid for an actual acoustic environment where the effect of room reverberation should be considered. A more realistic model for the microphone signals would be

$$x_i(n) = g_i * s(n) + w_i(n), \tag{2}$$

where $*$ denotes the convolution operation, and g_i is the channel impulse response function between the source and the i-th microphone. In the AED method, the impulse response

is assumed to be a finite impulse response (FIR) filter $\boldsymbol{g_i}$ with length $M(< N)$ as follows:

$$\boldsymbol{g_i} = \left[\, g_{i,0} \;\; g_{i,1} \;\; \cdots \;\; g_{i,M-1} \,\right]^T . \tag{3}$$

where T denotes the transpose.

For a noiseless case, $w_i(n) = 0$, the following relation

$$\mathbf{x}_1^T(n)\,\mathbf{g}_2 = \mathbf{x}_2^T(n)\,\mathbf{g}_1, \tag{4}$$

holds with

$$\boldsymbol{x}_i(n) = [x_i(n) \quad x_i(n-1) \quad \cdots \quad x_i(n-M+1)]^\tau \tag{5}$$

representing the vector of M signal samples arranged in a reverse order, since $x_1^* g_2 = (s^* g_1)^* g_2 = (s^* g_2)^* g_1 = x_2{}^* g_1$. If we define the $2M \times 2M$ dimensional covariance matrix $\mathbf{R}$ of vectors $x_1(n)$ and $x_2(n)$,

$$\mathbf{R} = \begin{bmatrix} E[\mathbf{x}_1(n)\mathbf{x}_1^T(n)] & E[\mathbf{x}_1(n)\mathbf{x}_2^T(n)] \\ E[\mathbf{x}_2(n)\mathbf{x}_1^T(n)] & E[\mathbf{x}_2(n)\mathbf{x}_2^T(n)] \end{bmatrix}, \tag{6}$$

together with the 2M $\times$ 1 dimensional vector u by concatenating the impulse response functions,

$$\mathbf{u} = \begin{bmatrix} \mathbf{g}_2 \\ -\mathbf{g}_1 \end{bmatrix}, \tag{7}$$

then Equations (4)–(7) immediately yield $\mathbf{Ru} = \mathbf{0}$, meaning that the vector $\mathbf{u}$ is in the null space of the covariance matrix. Equivalently, $\mathbf{u}$ is the eigenvector of $\mathbf{R}$ corresponding to the zero eigenvalue. Therefore, the two impulse responses ($\mathbf{g_1}$ and $\mathbf{g_2}$) can be found by determining this eigenvector. Then, the time delay estimate $\hat{\tau}$ is given by the time difference between their dominant peaks, or by the difference between the two direct paths, that is,

$$\hat{\tau} = \arg\max_k \left|g_{1,k}\right| - \arg\max_k \left|g_{2,k}\right| . \tag{8}$$

The existence of noise regularizes the covariance matrix, so $\mathbf{R}$ no longer has a zero eigenvalue. Noting that $\mathbf{R}$ is positive definite rather than positive semi-definite, the alternative is to find the smallest eigenvalue and its associated eigenvector. For the latter, Benesty has proposed an iterative way to minimize the estimation error:

$$e(n) = \frac{\mathbf{u}^T(n)\mathbf{x}(n)}{||\mathbf{u}||_2}, \tag{9}$$

where $||\cdot||$ is the l_2-norm of a vector, and $\mathbf{x}(\mathrm{n}) = \left[\boldsymbol{x}_1^T(n)\boldsymbol{x}_2^T(n)\right]^T$ is a $(2M \times 1)$-dimensional vector. Furthermore, an efficient adaptation (or update) rule was developed as follows:

$$\mathbf{u}(n+1) = \frac{\mathbf{u}(n) - \mu e(n)\mathbf{x}(n)}{||\mathbf{u}(n) - \mu e(n)\mathbf{x}(n)||_2}, \tag{10}$$

where μ (> 0) is the step size. Afterward, the estimation of eigenvector $\mathbf{u}$ is attainable after convergence.

Bear in mind that the goal of AED is not to estimate (even approximately) the impulse response functions, but rather to estimate the time delay by simply reading the index difference of their peaks. In this regard, a heuristic initialization for the elements $u_i(n)$ of vector $\mathbf{u}(n)$ at $n = 0$ plays a vital role in accomplishing this task. To allow a positive or negative relative delay, Equation (10) starts with a vector $\mathbf{u}(0)$ with only one nonzero element in the middle of its first half. That is,

$$u_k(0) = \begin{cases} 1 & \text{for } k = M/2 \\ 0 & \text{for } k \neq M/2, \end{cases} \tag{11}$$

the peak at $k = M/2$ is always dominant in comparison with other components in the first half of $\mathbf{u}$ throughout the iteration process. Such an invariant peak is related to the direct path of the impulse response $\mathbf{g_2}$. Another (negative) peak appearing in the last half of $\mathbf{u}$ would designate the direct path related to the impulse response $-\mathbf{g_1}$. Hence, it should be emphasized that the initial configuration of Equation (11) signifies the direct paths only, and in turn hinders a faithful description of the reverberation.

In addition, AED presumes a proper choice of the parameters μ and M, a factor which was not discussed in the previous studies. In treating the first item, it is worth mentioning that the update rule in Equation (10) was derived using the least-mean-square (LMS) algorithm, behaving just like the steepest-decent algorithm. It is well known that the stability of the steepest-decent method is governed by the eigenvalues of the covariance matrix $\mathbf{R}$ [30]. Accordingly, the step size μ needs to obey the following relationship to ensure the convergence of the algorithm:

$$0 < \mu < \frac{1}{\lambda_{\max}} \tag{12}$$

where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{R}$.

There seems to be no specific guideline for the filter length M. As long as the initialization scheme in Equation (11) is followed, however, the minimum requirement can be established as follows:

$$\frac{M}{2} > |\text{true delay}|, \tag{13}$$

explaining the appearance of the (negative) peak of $-\mathbf{g_1}$ in either side of $u_{M/2}(n)$. However, some trial and error is inevitable, as the true delay information is unavailable in general. A more practical approach is to begin with a sufficiently large M, and gradually reduce it until the estimated delay approaches a certain value.

By letting the true delay to 30 samples with reference to channel 2, some examples are presented in the upper row of Figure 1 for demonstration purposes. In an anechoic environment ($\mathrm{T}_{60} = 0.0$ s), the index difference between the peaks of two impulse responses is in exact agreement with the assumed delay. However, an increase in reverberation induces the appearance of spurious peaks in the estimations of $\mathbf{g_1}$, and eventually leads to anomalous delay estimates.

B. AUTOREGRESSIVE (AR) MODEL-BASED LINEAR PREDICTION PREFILTERING

The previous result implies that AED may fail when the room is subject to more reverberation. The results inevitably

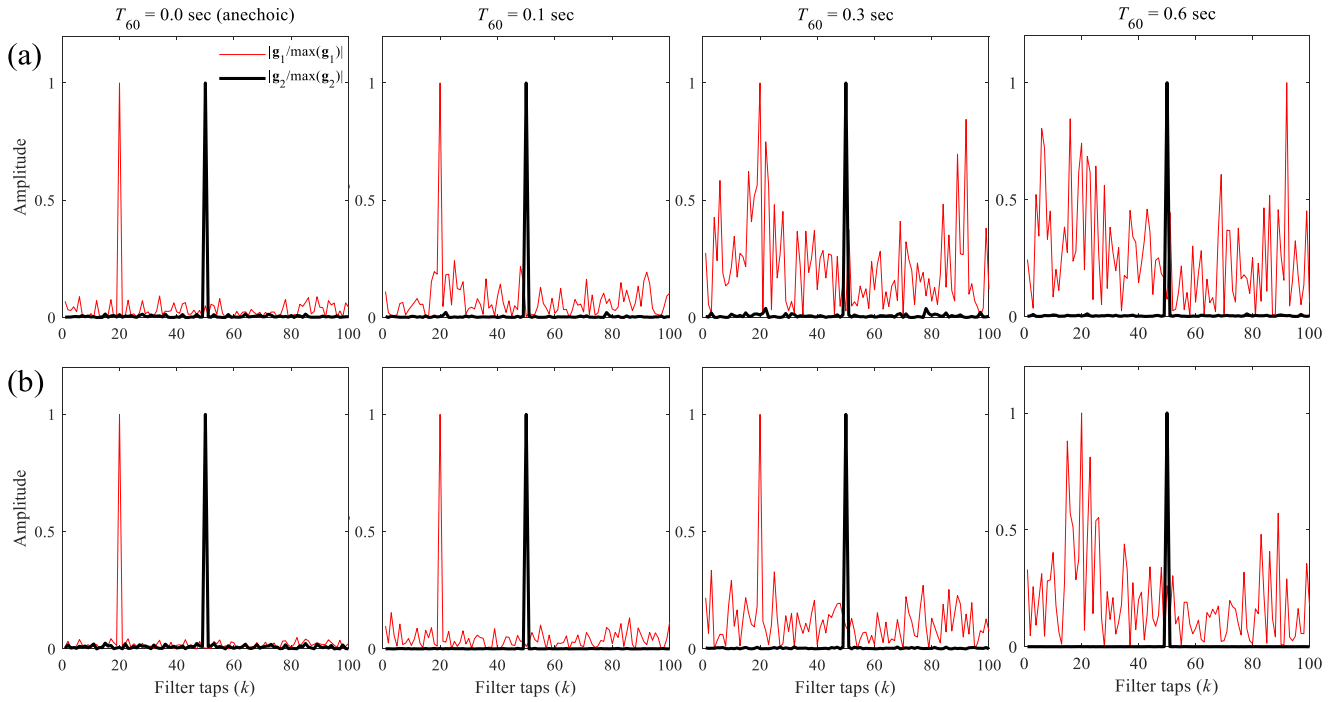

FIGURE 1. Examples of eigenvector u estimation: (a) (upper row) conventional AED, (b) (lower row) AED with AR prefiltering. Problem: Case A, SNR: 20 dB, True delay: 30 samples (Reference: microphone 2).

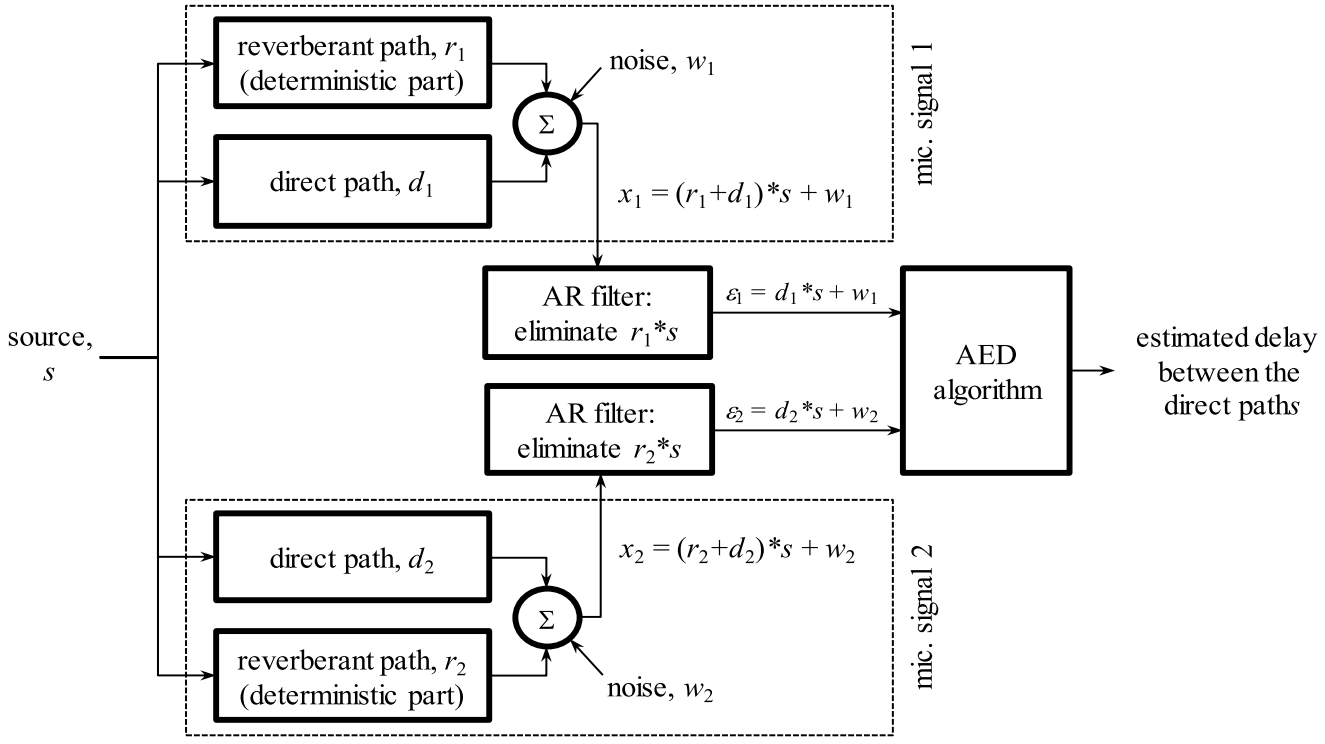

FIGURE 2. Conceptual diagram of AR model-based prefiltering.

become worse in the presence of strong noise. As mentioned previously, one possible explanation might be that AED is not concerned with precise modeling of the room impulse response. For example, Figure 1(a) confirms again that all the identified impulse responses have no consecutive ringing immediately after the direct path component.

Hence, it is desirable to remove the echoes from signals using an appropriate filter before performing AED. If we assume that the reverberation mainly comes from the resonant behavior of the room, it can be regarded a type of deterministic random process, which is described well by the AR model. Once the dereverberation is done, that is, the deterministic part is eliminated from the microphone signal by using the model as a linear prediction filter, the direct path component plus noise is left intact in the residual. Then, feeding the residual, rather than the original, into the AED algorithm would enhance the algorithm's TDE performance. The proposed concept is illustrated in Figure 2, where the AR model-based prefiltering is conducted for both microphone signals.

As stated, the linear prediction model employs the AR process in order to synthesize the deterministic (or predictable) part of a signal. More specifically, the i-th ($i = 1,2$) microphone signal is modeled as a weighted sum of p past values:

$$x_i(n) = -\sum_{j=1}^{p} a_{i,j} x_i(n-j) \quad + \varepsilon_i(n), \tag{14}$$

where the first part of the RHS denotes the AR part describing the deterministic component (produced by the reverberation), and the latter denotes the residual (direct path component with noise). $a_{i,j}$ denotes the model coefficients with $|a_{i,j}| < 1$ ($i = 1, 2$ and $j = 1, 2, \ldots, p$). These coefficients satisfy the Yule-Walker equation in the following:

$$\begin{aligned} r_{x_i x_i}(k) &= -\sum_{j=1}^{p} a_{i,j} r_{x_i x_i}(k-j) \quad \text{for } k = 1,\ 2,\ \ldots,\ p \\ r_{x_i x_i}(0) &= -\sum_{j=1}^{p} a_{i,j} r_{x_i x_i}(-j) + \sigma^2 \quad \text{for } k = 0, \end{aligned} \tag{15}$$

where σ^2 is the variance of the residual, and r_{xixi} is autocorrelation function whose estimate is given by

$$r_{x_i x_i}(k) = \frac{1}{N} \sum_{n=0}^{N-1} x_i(n) x_i(n-k). \tag{16}$$

The classical Gaussian elimination method is plausibly effective for solving the above matrix equation, but computationally inefficient. Instead, the recursive method proposed by Levinson–Durbin [31] is more common for the determination of $a_{i,j}$ and σ^2.

Upon the completion of AR parameters, Equation (14) is z-transformed to yield

$$X_i(z) = \frac{E_i(z)}{A_i(z)}, \tag{17}$$

where

$$\frac{1}{A_i(z)} = \frac{1}{1 + \sum_{j=1}^{p} a_{i,j} z^{-j}} \tag{18}$$

is often referred as the all-pole filter, because no zero appears in its numerator. Intuitively, from Equation (17), the output $X_i(z)$ is produced by the system transfer function $1/A_i(z)$ together with the input $E_i(z)$. Then, the discrete time version of the residual can be computed from the following convolution property:

$$\varepsilon_i(n) = Z^{-1}[A_i(z)] * x_i(n), \tag{19}$$

where Z^{-1} denotes the inverse z-transformation. As mentioned, we anticipate that in $\varepsilon_i(n)$, the direct path component can be recovered while reducing the strength of room reverberation as much as possible.

The most important step in AR modeling is the proper selection of model order p—that is, determining how many previous values of x_i should be taken into account. A model with an excessively small order yields a biased prediction that underfits the signal, whereas an excessively large order overfits the signal and fails to generalize the prediction. Depending on the application, a number of different criteria for resolving the optimum order have been devised [32]. If the residual signal can be assumed to be stationary white noise, as in the present study, using the Akaike information criterion (AIC) [33] is a standard approach. In this method, the following function is considered:

$$\mathrm{AIC}(p) = N \ln \sigma^2(p) + 2p. \tag{20}$$

The variance $\sigma^2(p)$ representing the approximation error is computed by Equation (15), and generally decreases with an increase of p. The term $2p$ was introduced to impose a penalty for selecting a high order. Thus, the optimum model order is chosen by minimizing AIC(p).

III. PERFORMANCE EVALUATION VIA NUMERICAL EXPERIMENTS

A. SETUP AND PROCEDURE

The purpose of these experiments was to validate the effectiveness of AR prefiltering in enhancing AED. Hence, reverberant environments with room dimensions of $L_x = 3.0$ m, $L_y = 4.0$ m, and $L_z = 2.5$ m were numerically simulated by employing the image source method [34], [35]. According to Eyring's formula [36], the acoustic property of a room with volume V and total surface area S is characterized by the reverberation time, viz.,

$$T_{60} = \frac{0.163V}{-S \ln(1-\gamma)}, \tag{21}$$

where the flat surfaces (walls, ceiling, and floor) are assumed to possess a uniform absorption coefficient γ without any dependence on the frequency and/or the incident angle. The key parameters for the simulation were as follows:

- Reverberation time, T_{60}: varied from 0.0 s (anechoic condition) to 0.6 s, step size 0.1 s (total 7 steps)
- Signal-to-noise ratio (SNR): varied from 20 dB to 0 dB (0, 5, 10, 20 dB, total 4 steps)
- Speed of sound: 343 m/s
- Sensor positions: microphone 1 at (x, y, z) = (1.00, 2.00, 1.43) m and microphone 2 at (2.30, 1.20, 1.00) m

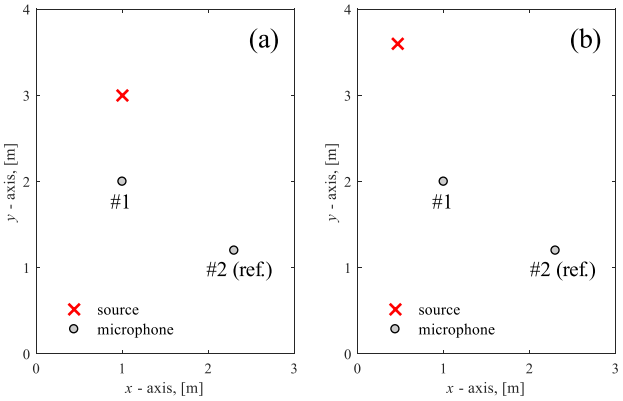

FIGURE 3. Arrangement of two microphones and the source in a room with $L_x = 3.0$ m, $L_y = 4.0$ m and $L_z = 2.5$: (a) Case A, (b) Case B. For the sampling frequency of 8 kHz, the true delay was 30 and 32 samples for Cases A and B, respectively.

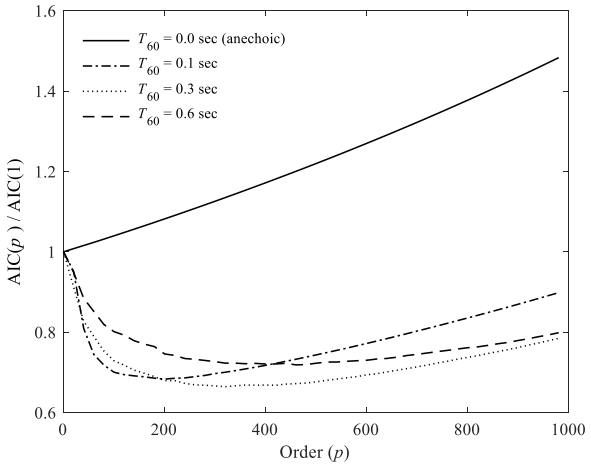

FIGURE 4. Evaluation of Akaike information criterion [33] as a function of the AR model order p. The results are for the impulse response between the source and microphone 1 in Case A (SNR = 20 dB).

- Source position: omnidirectional source at (1.001, 3.00, 1.60) m for Case A, and at (0.565, 3.50, 1.60) m for Case B
- True delay between direct paths with reference to microphone 2: +30.0 samples for Case A, +32.0 samples for Case B

For generalization purposes, the simulation was subjected to variations in the reverberation as well as in the noise level. The SNR is defined as follows:

$$\mathrm{SNR} = 10\log_{10}\left(\sigma_x^2/\sigma_n^2\right), \tag{22}$$

where σ_x^2 is the variance of an uncontaminated signal and σ_n^2 is the additive noise. Two cases of source positions were considered, as illustrated in Figure 3. The latter (Case B) posed a more difficult problem, as the source was moved to the corner of the room. Note that these source positions were finely adjusted to ensure the assumed true delay was an integral multiple of the signal samples for the specified sampling frequency 8 kHz, thus avoiding the necessity of further treatment dealing with resolution issues (for instance, interpolation between the samples).

For the given reverberation time, room impulse responses between the source and two microphones were computed by Lehmann's MATLAB code [37], an improved version of Allen and Berkley's implementation [34] of the image source method. A white Gaussian source was convolved with the simulated impulse responses to generate microphone signals. Finally, mutually independent random noise was superimposed on the clean signals to vary the SNR. The length of the used signals was $N = 5000$ samples (= 0.625 s) for all experiments.

Parameter selection scheme for the AED was performed according to the method explained in Section 2. The step size μ was automatically determined by setting $\mu = 1/(10\lambda_{\max})$, which is sufficiently smaller than the allowable limit in Equation (12). Besides, the length of the adaptive eigenvector u

TABLE 1. Optimum Orders of AR model (Case A, SNR = 20 dB).

T_{60} [s]	Optimum order, p	
	Channel 1 (Source - microphone 1)	Channel 2 (Source - microphone 2)
0.0	1	1
0.1	197	141
0.3	306	278
0.6	467	334

was determined as $L = 2M = 200$ according to the relation in (13). This made the number of iterations (N - M) excessively large. Fortunately, the AED method converges very quickly. Therefore, even though the full number of iterations was not completed, the algorithm was terminated once the normalized error defined in Equation (9) dropped below the specified tolerance, i.e., 1×10^{-3}.

B. RESULTS: EFFECTS OF REVERBERATION AND NOISE

The outputs of Case A are presented first. Some of these results were already observed in Figure 1(a), wherein the delay estimate using the conventional AED deteriorates according to the increase in reverberation. As the earliest step for applying AR prefiltering, the optimum order of the AR model is of primary interest. Thus, AIC was evaluated as a function of the order p for different levels of the reverberation time T_{60}. As shown in Figure 4 and summarized in Table 1, the optimum order where AIC is minimized generally increases along with T_{60}. Fewer poles (past values of $x_i(n)$) would be necessary to describe the impulse response in a shorter room, whereas a longer reverberant decay would require a higher order for approximation. Therefore, it is concluded that the model orders for different levels of T_{60} have been identified reasonably well.

Figure 1(b) shows the eigenvector $\mathbf{u}$ estimation with the use of AR prefiltering, in which the trend is similar to that of the case without AR prefiltering. That is, the method is

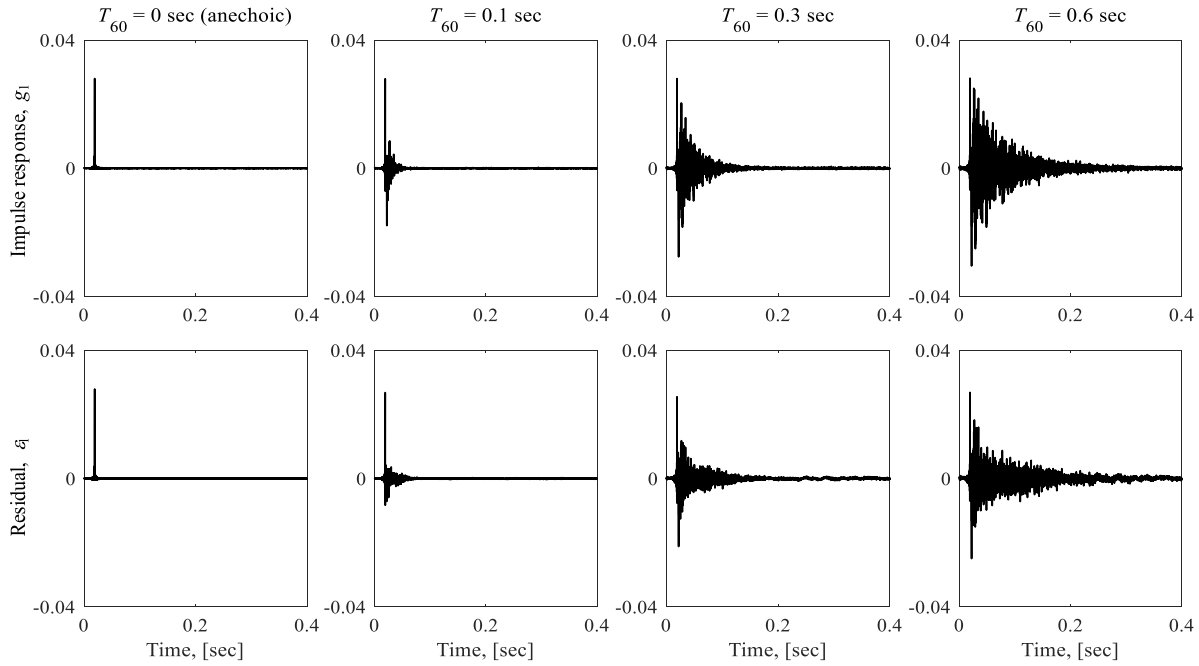

FIGURE 5. Microphone signals for impulse excitation (upper row) and the residual after AR prefiltering (lower row). The results are for microphone signal 1 in Case A (SNR = 20 dB).

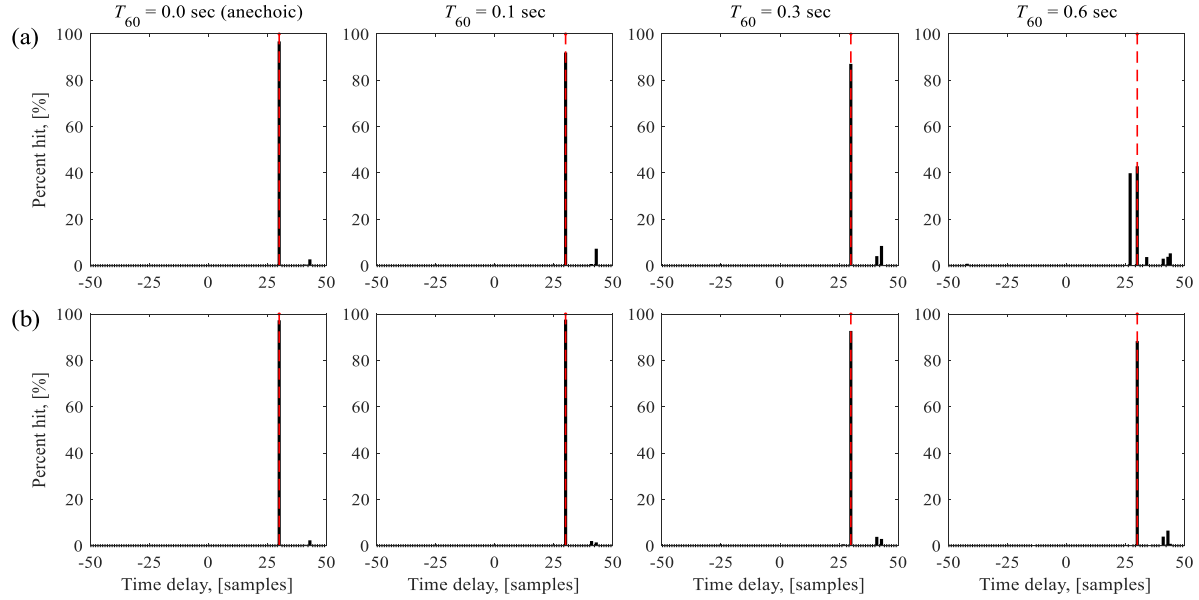

FIGURE 6. Histograms of TDE (a) without and (b) with AR prefiltering. The dotted red line denotes the true delay. The results are for Case A (SNR: 20 dB).

still unable to prevent the growth of erroneous peaks along with T_{60}, and even may fail for very strong reverberation. However, as the contrast of the main peak becomes much clearer, the application of prefiltering significantly enhances the performance of AED. These examples can be regarded as direct evidence supporting the usefulness of the proposed method.

Nevertheless, it would be appropriate to study the effect of reverberation in more detail. If the AR prefiltering perfectly removes the reverberation, the residual should be identical to the source signal except for a time shift. For an impulse excitation, for which we obtain the room impulse response, the delta function is expected as an ideal residual. In other words, the effectiveness of AR filtering can be assessed in

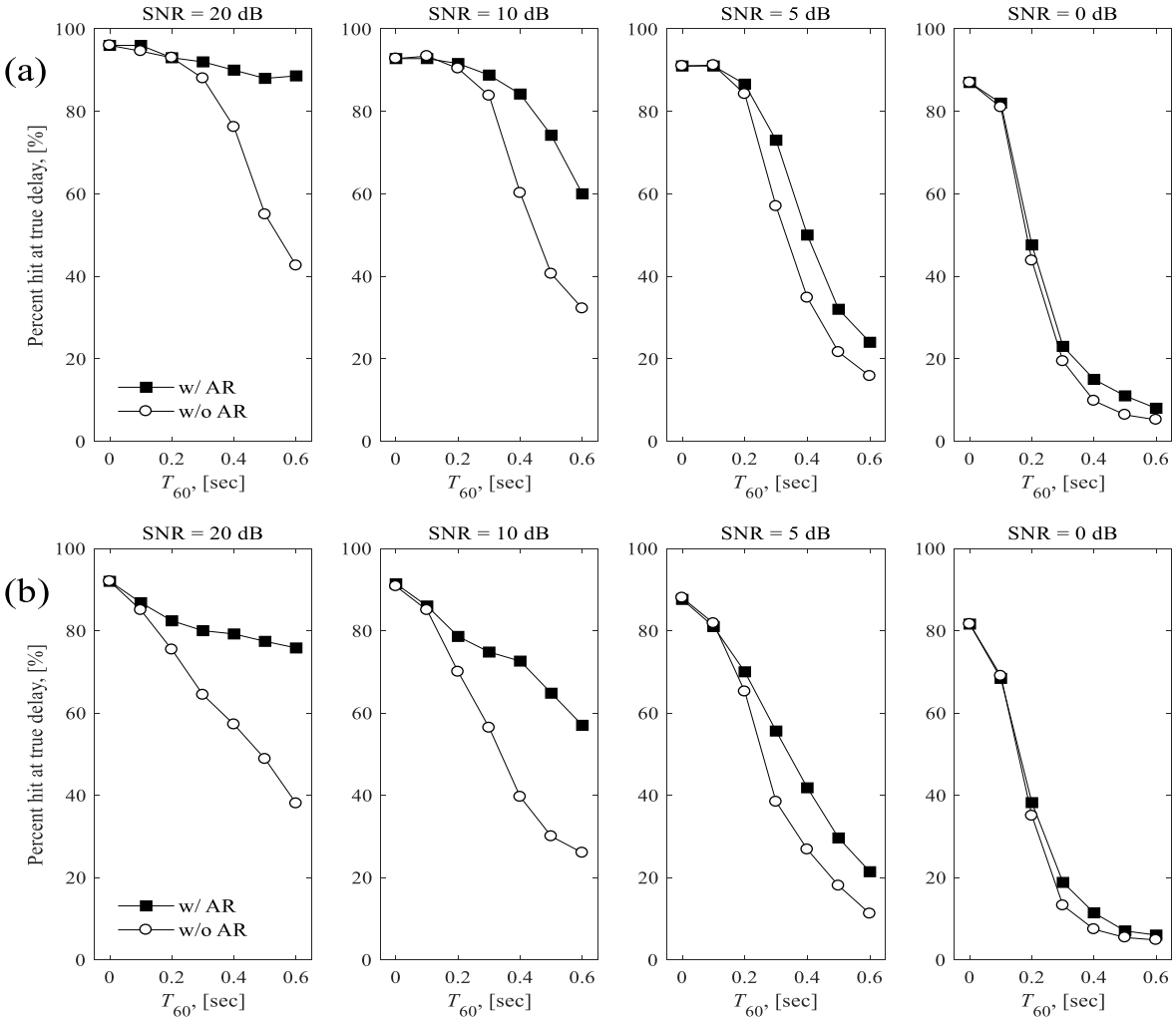

FIGURE 7. Comparison of percent hits at the true delay for various levels of T_{60} and SNR: (a) Case A, (b) Case B.

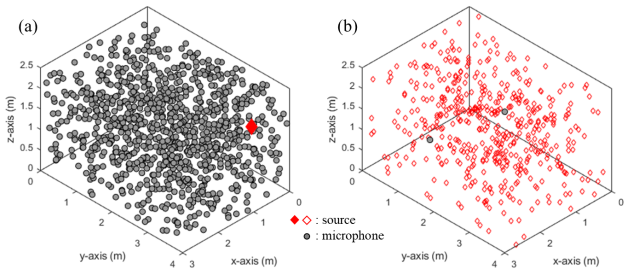

FIGURE 8. Arrangement of two microphones and the source in a room, L_x = 3.0 m, L_y = 4.0 m and L_z = 2.5: (a) Case C, (b) Case D. Sampling frequency is 8 kHz; source location is fixed in (a) and sensor locations are fixed in (b).

terms of how close the residual obtained for an impulse source is to the ideal delta function. To check this, the response for an impulse excitation (i.e., the impulse response) and the corresponding residuals were investigated. Figure 5 shows that AR filtering for a less echoic environment achieves a closer approximation to the original excitation. However, an increase in T_{60} accentuates the reverberant energy in the residual. Such deviations from the ideal delta function imply the degradation of TDE performance in the subsequent AED analysis.

To perform further validation, Monte Carlo simulations of 1000 independent trials were carried out for each test condition; here, reverberation from 0 (anechoic) to 0.6 s was considered as the key variable under the assumption of 20 dB SNR. Figures 6(a) and (b) show histograms of the delay estimates without and with AR prefiltering, respectively. The true delay is depicted with a dotted line. The conventional AED method suffers from poor estimation capability for a high reverberation level, whereas the AR filtering relieves such degradation remarkable well. When $T_{60} = 0.6$ s, for example, the prefiltering enhanced the delay estimation performance by approximately 50%.

In the above, the impact of reverberation on the time delay estimation was investigated, for which a high SNR was assumed. However, we need to deal with not only the

effects of reverberation, but also the noise. Thus, the second experiment also involved a set of data obtained in noisy environments, but gathers the percent hits at the true delay rather than the histogram. In summary, there were a total of 56,000 numerical experiments according to various SNR values and amounts of reverberation. As illustrated in Figure 7(a), it is clear that lowering the SNR degrades the estimation performance. For the worst case wherein very strong reverberation and noise prevail, the improvement obtained from AR prefiltering does not exceed a few percent. However, such severity is rarely encountered in practical applications such as teleconferencing and robot audition systems. When the noise level is moderate, the method is considered to be valid even in a highly echoic room.

Experiments using the same variations of T_{60} and SNR were conducted for Case B (Figure 7(b)). As the source moves toward the corner of the room, the reverberation structure becomes more complex, eventually making it harder to identify them. Compared to Case A in Figure 7(a), the results are worse. However, improvement is still attainable, and the trends are consistent with those of the previous case. This verifies that the prefiltering treatment is not substantially influenced by the source position.

C. RESULTS: EFFECTS OF SOURCE AND MICROPHONE LOCATIONS

Additional experiments were performed to more clearly confirm the effectiveness of the proposed prefiltering methodology. The performance of the proposed method was reviewed by changing the position of the source while the positions of the sensors were fixed, and by changing the positions of the sensors while the position of the source was fixed. The scenarios in which performance was examined are summarized as follows:

Case C:

- Reverberation time, T_{60}: 0.6 s
- Signal-to-noise ratio (SNR): 20 dB
- Sensor positions (microphones 1 & 2): random generation (500 trails)
- Source position (omnidirectional source): (1.001, 3.00, 1.60) m, same as in Case A

Case D:

- Reverberation time, T_{60}: 0.6 s
- Signal-to-noise ratio (SNR): 20 dB
- Sensor positions: microphone 1 at $(x, y, z) = (1.00, 2.00, 1.43)$ m and microphone 2 at (2.30, 1.20, 1.00) m
- Source position (omnidirectional source): random generation (500 trails)

The sensor and source locations for the generated scenarios are shown in Figures 8(a) and (b). The performance improvement of the time delay estimation from the simulations is summarized in Table 2; in these trials, it was confirmed that the performance improved by 118% and 48.6% when the proposed method was used as the prefiltering technique. In particular, it was confirmed that the performance of AED, which is sensitive to the position of the sensors, was remarkably improved by a factor of more than 2.

TABLE 2. Effects of Source and Microphone Locations (Cases C & D).

Item	Case Study	
	Case C	Case D
Source	(1.001, 3.00, 1.60) m	Random: #500
Microphone	#1 (Random: #500) #2 (Random: #500)	#1 (1.00, 2.00, 1.43) m #2 (2.30, 1.20, 1.00) m
#Success / #500	Performance of Time Delay Estimation (Success: Estimated delay = True delay)	
(a) w/ AR	28.8 %	54.4%
(b) w/o AR	13.2 %	36.6%
Ratio ((a)/(b))	2.18	1.49

IV. CONCLUSION

The main topic of this study was the improvement of TDE performance in a reverberant environment, using the AED algorithm as an example. AR-model based prefiltering was proposed to eliminate the reverberation in the received signal, leaving the direct path component in the residual. Then, the results were fed into the AED algorithm to suppress the appearance of erroneous peaks in the estimated impulse responses, and subsequently to sharpen the peaks that were related to the true delay. Monte Carlo-based numerical experiments were conducted to verify that the prefiltering treatment effectively improves TDE performance. Provided that the noise level is moderate, the method's validity is established even in a highly echoic room. From 1,000 simulations, it was confirmed that the existing AED method has a low probability of 42.9% when there is a relatively high amount of echo, whereas the proposed method can find the source location with an accuracy of 88.3%. Furthermore, the outcomes of additional numerical experiments demonstrated that the performance improvement is largely independent from the source and sensor locations. Therefore, the proposed method can be applied to enhance the performance of the conventional AED method.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer, 2008, doi: 10.1007/978-3-540-78612-2.

[2] I. J. Tashev, *Sound Capture and Processing: Practical Approaches*. Hoboken, NJ, USA: Wiley, 2009.

[3] R. R. Fay, *Sound Source Localization*, vol. 25. New York, NY, USA: Springer, 2006.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976, doi: 10.1109/TASSP.1976.1162830.

[5] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Dept. Division Eng., Brown Univ., Providence, RI, USA, 2000.

[6] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, Jan. 2011.

[7] H. He, X. Wang, Y. Zhou, and T. Yang, "A steered response power approach with trade-off prewhitening for acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 143, no. 2, pp. 1003–1007, Feb. 2018.

[8] T. Padois, "Acoustic source localization based on the generalized cross-correlation and the generalized mean with few microphones," *J. Acoust. Soc. Amer.*, vol. 143, no. 5, pp. EL393–EL398, May 2018.
[9] B. Yang, H. Liu, C. Pang, and X. Li, "Multiple sound source counting and localization based on TF-wise spatial spectrum clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1241–1255, Aug. 2019.
[10] H. Liu, B. Yang, and C. Pang, "Multiple sound source localization based on TDOA clustering and multi-path matching pursuit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 3241–3245.
[11] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, 2000, doi: 10.1121/1.428310.
[12] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, Nov. 2003.
[13] X. Alameda-Pineda and R. Horaud, "Geometrically-constrained robust time delay estimation using non-coplanar microphone arrays," in *Proc. EUSIPCO*, Aug. 2012, pp. 1309–1313.
[14] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sep. 2004.
[15] T. Padois, O. Doutres, and F. Sgard, "On the use of modified phase transform weighting functions for acoustic imaging with the generalized cross correlation," *J. Acoust. Soc. Amer.*, vol. 145, no. 3, pp. 1546–1555, Mar. 2019.
[16] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Process.*, vol. 85, no. 1, pp. 177–204, 2005.
[17] M. Cobos, F. Antonacci, L. Comanducci, and A. Sarti, "Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1270–1281, 2020.
[18] H. He, J. Chen, J. Benesty, W. Zhang, and T. Yang, "A class of multichannel sparse linear prediction algorithms for time delay estimation of speech sources," *Signal Process.*, vol. 169, Apr. 2020, Art. no. 107395.
[19] R. Badeau, G. Richard, and B. David, "Sliding window adaptive SVD algorithms," *IEEE Trans. Signal Process.*, vol. 52, no. 1, pp. 1–10, Jan. 2004, doi: 10.1109/TSP.2003.820069.
[20] T. Chonavel, B. Champagne, and C. Riou, "Fast adaptive eigenvalue decomposition: A maximum likelihood approach," *Signal Process.*, vol. 83, no. 2, pp. 307–324, 2003, doi: 10.1016/s0165-1684(02)00417-6.
[21] R. Wang, F. Gao, M. Yao, and H. Zou, "Low complexity adaptive algorithm for generalized eigenvalue decomposition," in *Proc. 8th CHINACOM*, Aug. 2013, pp. 690–693, doi: 10.1109/chinacom.2013.6694681.
[22] V. Bhat, I. Sengupta, and A. Das, "An adaptive audio watermarking based on the singular value decomposition in the wavelet domain," *Digit. Signal Process.*, vol. 20, no. 6, pp. 1547–1558, 2010, doi: 10.1016/j.dsp.2010.02.006.
[23] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Adv. Signal Process.*, vol. 2006, no. 1, pp. 1–19, Dec. 2006, doi: 10.1155/ASP/2006/26503.
[24] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—A tutorial," *Mech. Syst. Signal Process.*, vol. 25, no. 2, pp. 485–520, 2011, doi: 10.1016/J.YMSSP.2010.07.017.
[25] N. Sawalhi, R. B. Randall, and H. Endo, "The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis," *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2616–2633, 2007, doi: 10.1016/j.ymssp.2006.12.002.
[26] H. Endo and R. B. Randall, "Enhancement of autoregressive model based gear tooth fault detection technique by the use of minimum entropy deconvolution filter," *Mech. Syst. Signal Process.*, vol. 21, no. 2, pp. 906–919, 2007, doi: 10.1016/j.ymssp.2006.02.005.
[27] W. Wang and A. K. Wong, "Autoregressive model-based gear fault diagnosis," *J. Vib. Acoust.*, vol. 124, no. 2, pp. 172–179, 2002, doi: 10.1115/1.1456905.
[28] J. Mourjopoulos and M. A. Paraskevas, "Pole and zero modeling of room transfer functions," *J. Sound Vib.*, vol. 146, no. 2, pp. 281–302, 1991, doi: 10.1016/0022-460X(91)90764-B.
[29] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 320–328, Apr. 1994, doi: 10.1109/89.279281.
[30] A. Poularikas, *Adaptive Filtering: Fundamentals of Least Mean Squares With MATLAB*. Boca Raton, FL, USA: CRC Press, 2015.
[31] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.
[32] R. Palaniappan, "Towards optimal model order selection for autoregressive spectral analysis of mental tasks using genetic algorithm," *Int. J. Comput. Sci. Netw. Secur.*, vol. 6, pp. 153–162, Jan. 2006.
[33] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974, doi: 10.1109/TAC.1974.1100705.
[34] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979, doi: 10.1121/1.382599.
[35] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, 2008, doi: 10.1121/1.2936367.
[36] I. L. Vér and L. L. Beranek, *Noise and Vibration Control Engineering: Principles and Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[37] E. A. Lehmann. (2016). *Image-Source Method: MATLAB Code Implementation*. Accessed: Mar. 1, 2019. [Online]. Available: http://www.eric-lehmann.com/

YUN-HO SHIN received the M.S. and Ph.D. degrees in mechanical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2004 and 2009, respectively. From 2010 to 2019, he was a Senior Researcher with the System Dynamic Research Department, Korea Institute of Machinery and Materials, Daejeon. Since 2019, he has been an Assistant Professor with the Department of Safety Engineering, Chungbuk National University, Cheongju, South Korea. His research interests include passive and active vibration isolation, nonlinear control technologies and their applications, and naval ship survivability.

JEUNG-HOON LEE received the M.S. and Ph.D. degrees in mechanical engineering from KAIST, Daejeon, South Korea, in 2002 and 2007, respectively. After nine years of industrial experience in SSMB of Samsung Heavy Industries Company Ltd., in 2016, he joined the Department of Mechanical Engineering, Changwon National University, Changwon, South Korea, as an Associate Professor. His research interests include several fields, such as passive vibration isolation and acoustic cavitation.