Exploring conventional enhancement and separation methods for multi-speech enhancement in indoor environments

Yangjie Wei | Ke Zhang | Dan Wu | Zhongqi Hu

College of Computer Science and Engineering, Northeastern University, Shenyang, China

Correspondence
Yangjie Wei, College of Computer Science and Engineering, Northeastern University, Shenyang, China.
Email: weiyangjie@cse.neu.edu.cn

Funding information
The Natural Science Foundation of China, Grant/ Award Number: 61973059

Abstract
Speech enhancement is an important preprocessing step in a wide diversity of practical fields related to speech signals, and many signal-processing methods have already been proposed for speech enhancement. However, the lack of a comprehensive and quantitative evaluation of enhancement performance for multi-speech makes it difficult to choose an appropriate enhancement method for a multi-speech application. This work aims to study the implementation of several enhancement methods for multi-speech enhancement in indoor environments of $T_{60} = 0$ s and $T_{60} = 0.3$ s. Two types of enhancement approaches are proposed and compared. The first type is the basic enhancement methods, including delay-and-sum beamforming (DSB), minimum variance distortionless response (MVDR), linearly constrained minimum variance (LCMV), and independent component analysis (ICA). The second type is the robust enhancement methods, including improved MVDR and LCMV realized by eigendecomposition and diagonal loading. In addition, online enhancement performance based on the iteration of single-frame speech signals is researched, as is the comprehensive performance of various enhancement methods. The experimental results show that the enhancement effects of LCMV and ICA are relatively more stable in the case of basic enhancement methods; in the case of the improved enhancement algorithms, methods that employ diagonal loading iterations show better performance. In terms of online enhancement, DSB with frequency masking (FM) yields the best performance on the signal-to-interference ratio (SIR) and can suppress interference. The comprehensive performance test showed that LCMV and ICA yielded the best effects when there was no reverberation, while DSB with FM yielded the best SIR value when reverberation was present.

1 | INTRODUCTION

Speech enhancement refers to the extraction of useful speech signals from noisy backgrounds and the suppression or reduction of noise interference when speech signals are disturbed or even submerged by various types of noise. In general, all techniques that can extract pure speech signals belong to the category of speech enhancement, which has been used extensively in voice calls, teleconferences, military eavesdropping, and speech recognition and become a useful front-end processing module in many speech-related systems.

Speech enhancement can be subclassified into signal-processing-based and machine-learning-based enhancements. According to the number of microphones used for speech reception, the former enhancement can be further subclassified into single-channel and multi-speech enhancements (also known as microphone-array speech enhancements) [1]. Compared with single-channel speech enhancement, enhancement methods based on a microphone array can realize spatial filtering because the locations of all sound sources are considered. Additionally, these methods can suppress directional noise. With the rapid development of artificial intelligence and other related technologies in recent years, many multi-speech-enhancement methods have been proposed to adapt to various practical scenarios [2–5]. However, in actual, complicated acoustic environments, various interference and aliasing phenomena could exist, and it is thus difficult to choose an appropriate speech-enhancement method for some...
practical tasks, especially in a multi-speech environment. The fundamental reasons for this are as follows: (1) Although the applicable area of each enhancement method can be defined according to its basic principles, there must have an unpredictable distance between the theoretical requirements and the specific scenarios because of the variations and coupling of actual acoustic features. (2) To comprehensively evaluate the performance of a multi-speech-enhancement method, a variety of indicators, such as the noise ratio, stability, degrees of noise suppression, and target loss, should be considered. However, few research studies have simultaneously calculated and compared these performance indicators in different acoustic environments. These problems have seriously affected the effective use of current speech-enhancement methods in real applications.

This work aims to study the performance of different enhancement methods for multi-speech in indoor environments of T60 = 0 s and T60 = 0.3 s. In contrast with most of the research focusing on performance evaluation, we first calculated and compared the source-to-artefacts ratio (SAR), signal distortion ratio (SDR), and signal-to-interference ratio (SIR) of the basic and adaptive enhancement methods for the offline aspect and then researched the online enhancement performance based on the iteration of single-frame speech signals and the comprehensive performance of various enhancement methods. Another differential aspect of our approach is selecting two specific categories of environments as targets, T60 = 0 s and T60 = 0.3 s, that emulate the acoustic scenes of most actual environments. Therefore, in this study, various performance evaluation indicators associated with extensively used speech-enhancement algorithms were calculated in several indoor environments, and the adaptability, limitations, and future improvement directions of these speech-enhancement methods were then analysed through comparisons.

In short, the first difference between this paper and existing results is that six speech-enhancement methods that belong to two types of approaches, beamforming and speech separation, were all denoted as a separation-matrix-related function, and then their enhancement performance evaluation was transformed into a precision evaluation of the separation matrix. Therefore, it is possible to test and compare these enhancement methods theoretically in the same indoor environments. The second difference is that to comprehensively evaluate the performance of these multi-speech-enhancement methods, a variety of quantitative indicators were introduced. Furthermore, except for the offline enhancement, the online enhancement, as well as the comprehensive enhancement, were all conducted. Therefore, the result in this study is considered useful for theoretical research and in actual applications of speech enhancement.

The rest of this paper is organized as follows: Section 2 introduces the principles of speech-enhancement methods based on microphone arrays. Section 3 describes the evaluation indicators of the performance of speech enhancement and Section 4 contains the experimental setup and the experimental results for the tasks of speech enhancement, and then comparison and discussion are given. Finally, Section 5 discusses the conclusions of this work.

2 | SPEECH-ENHANCEMENT METHODS BASED ON MICROPHONE ARRAYS

Beamforming and speech separation could both be used to enhance a speech signal [1]. The basic idea of beamforming is to obtain the final target speech by weighting the filter coefficient for each channel of a microphone array after the received signals have been aligned in the azimuth of the target sound source. Therefore, it can be divided into two steps: time alignment and weighting summation. According to different methods of setting weighting coefficients, beamforming can be further classified into two types, fixed beamforming and adaptive beamforming. In contrast with beamforming, speech separation can simultaneously estimate the speech components of all sound sources received by a microphone array. If more than one speech signal must be enhanced concurrently, speech separation has obvious advantages because it only needs one calculation time. Therefore, speech separation could be thought of as a more powerful speech-enhancement method.

2.1 | Fixed beamforming

Fixed beamforming refers to beamforming with fixed weighting coefficients for each channel; that is, the weighting coefficient of a particular sound source is constant and not affected by the characteristics of different speech or environments. However, the coefficients are different in different directions, and therefore, fixed beaming could focus on a sound source in an arbitrary direction by varying the weighting coefficients [6].

The simplest way to implement fixed beamforming is delay-and-sum beamforming (DSB). Its mathematical expression is as follows:

$$y(t) = \sum_{i=0}^{M-1} a_i x_i(t + \tau_i)$$  \hspace{1cm} (1)

where $x_i$ denotes the received signal of the $i^{th}$ microphone in the time domain; $\tau_i$ denotes the time-delay compensation for the $i^{th}$ microphone channel; $a_i$ denotes the weighting coefficient of the $i^{th}$ microphone channel; $M$ denotes the number of microphone channels; and $t$ denotes time index.

In practical applications, power levels are unevenly distributed in the frequency domain owing to the sparsity of speech signals, and the components of an interference signal could be dominant in some time–frequency bins whereby the power distribution of the target source may be neglected. The idea of frequency masking was introduced to solve this
problem. Its principle is based on the separation of $N$ sources with fixed beamforming once the positions of the target source and interference sources are known.

A binary mask in the frequency domain is defined as the following equation,

$$U_o(f) = \begin{cases} 1 & o = \arg \max_A |O_o(f)| \quad a = \theta_1 \cdots \theta_N \\ 0 & \text{others} \end{cases}$$

(2)

where $A$ represents the angle collection of the sound sources; $a$ is the member of $A$; $\theta_i$ is the angle of source $i$; $O$ represents the beam pattern; and $f$ represents frequency.

It can be observed from Equation (2) that when the beam directed to $o$ possesses the highest beam power, the corresponding mask value of the sound source $o$ is equal to unity, and the mask value of the other sources is equal to zero. Therefore, the results of beamforming with the use of a binary mask indicate that the sound source has maximum power in the gathering direction, while the power in the other directions is reduced. Because of the sparsity of speech signals, only one sound source occupies a time–frequency bin under most situations. When the beamforming focuses on the direction of a sound source in this time–frequency bin, the beamforming outcome will have greater power than those in the other directions. Therefore, shielding the frequency bins whose power is higher than the target source is an effective way to shield the components of the interference sources.

In addition to the previous fixed beamforming, differential microphone arrays, superdirective microphone arrays, and frequency-invariant beamformers are also fixed beamforming technologies. Fixed beamforming has the advantages of easy implementation and low computational complexity. DSB is the optimal filter in the case when the current environment only contains uncorrelated noise between channels. However, it has a poor suppression effect and environmental adaptability for directional interference signals (which belong to coherent noise sources). In addition, the effect of fixed beamforming is closely related to the number of sound sources: (1) When only one source exists, it can be demonstrated with the help of time-delay compensation that the phases of all the frequency components of the source are the same on different channels, and the power increases linearly after superposition. By contrast, the noise signals on different channels cancel each other. Therefore, speech enhancement of the sound source can be realized. (2) When both primary and secondary sources exist, the phases on different channels are not the same after time-delay compensation, and the superposition power of these sources cancel each other (the phase difference and the power cancelation in the phase range of $[-\pi, +\pi]$ are proportional to each other). Therefore, the enhancement effect of two sound sources is limited with fixed beamforming.

In fact, the signal after beamforming can be written as a unified form,

$$y = w^H x$$

(3)

where $y$ denotes the output signal after beamforming; $w$ denotes the weighting vector, which can also be considered as a filter; $x$ denotes the signal received by a microphone array; and $H$ denotes the conjugate transpose.

Therefore, in speech enhancement, the key problem is to choose the parameters for the weighting vector $w$. The enhancement methods obtain an appropriate $w$ with different constraints and different criteria because different enhancement methods have different theories to obtain $w$, and their results may be different even when the same source data is used. In the following section, some adaptive methods are introduced to explain the definition of $w$.

### 2.2 Adaptive beamforming

The basic idea of adaptive beamforming is to change the weighting coefficients for the purpose of retaining the desired speech and filtering out interference. There are two basic adaptive beamforming methods widely used for speech enhancement.

#### 2.2.1 Minimum variance distortionless response

Minimum variance distortionless response (MVDR) is to select the weighting coefficients of a beamformer with the purpose of minimizing the output power under the constraint that the target speech is not affected.

Then, the MVDR criterion can be expressed as

$$\min_w w^H R w \quad \text{subject to } w^H d = 1$$

(4)

where $R$ denotes the data covariance matrix of interference and noise and $d$ denotes the steering vector on the target direction.

Equation (4) shows that the optimization objective of MVDR is to minimize the beamforming power, and the constraint condition is that the filter coefficients are weighted in the direction of the target signal. The summation is 1, which ensures that the target source signal remains unchanged. The optimization problem can be directly solved by the Lagrange multiplier method, given by

$$w = R^{-1} d$$

(5)

#### 2.2.2 Linearly constrained minimum variance

Linearly constrained minimum variance (LCMV) is to minimize the output power of the microphone array keeping the gain of the useful signals within the desired scope.

Firstly, a group of linear constraints is defined by a constraint matrix $C$, where the column vectors are linearly independent, given by
\[ C^H w = g \] (6)

where \( g \) is called a constraint value vector.

The LCMV beamformer is constrained by Equation (6) to minimize the output power of interference and noises:

\[ \min_w w^H R w, \text{subject to } C^H w = g \] (7)

The optimization problem in Equation (7) can also be solved by the Lagrange multiplier method. The optimal weighting vector is given as

\[ w = R^{-1} C (C^H R^{-1} C)^{-1} g \] (8)

Typical constraints include distortion-free, directional, zero, and derivative constraints.

2.2.3 Improvement of adaptive beamforming

It can be seen from Equation (8) that the inverse of the covariance matrix \( R \) needs to be calculated while solving the weighting vector. If the covariance matrix \( R \) is irreversible, the solution of Equation (7) leads to distortion in enhanced signals. In theory, \( R \) must be irreversible if the actual number of sound sources is less than the number of microphone channels without interference. To solve this problem, scholars have proposed various robust adaptive methods. Generally, these methods can be divided into two types, eigenspace beamformer (ESB) [7] and diagonal loading [8].

(a) Eigenspace beamformer

Firstly, the received signal \( x \) is decomposed into a signal subspace and a noise subspace. Assuming that the number of signal sources is less than or equal to the number of microphone channels, \( R \) is decomposed into eigenvalues as

\[ R = \sum_{i=1}^{M} \lambda_i e_i e_i^H = e_1 \lambda_1 e_1^H + \cdots + e_n \lambda_n e_n^H \] (9)

where \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{P+1} = \cdots = \lambda_M = \sigma_n^2 \) are the eigenvalues of \( R \); \( e_i \) is the eigenvector corresponding to eigenvalue \( \lambda_i \); \( E_s = [e_1, e_2, \ldots, e_{P+1}] \) and \( E_n = [e_{P+2}, e_{P+3}, \ldots, e_M] \) are orthogonal to each other and they correspond to the signal subspace and the noise subspace, respectively; \( \Lambda_s = \text{diag} \{ \lambda_1, \lambda_2, \ldots, \lambda_P \} \); \( \Lambda_n = \sigma_n^2 I_{N-P} \) where \( N \) and \( P \) are sensor space dimension and signal subspace dimension, respectively.

Then, the optimal weighting vector based on the feature subspace is given as

\[ w = \frac{E_s \Lambda_s^{-1} E_s^H d}{d^H E_s \Lambda_s^{-1} E_s^H d} \] (10)

Equation (10) is usually solved using principle component analysis (PCA) to decompose and reduce dimensionality. The dimensionality after reduction could be equal to the number of sound sources. Of course, the number of sound sources is required to be known, and the enhancement performance will be affected by the frequency number of the speech signals. Therefore, adaptive eigen-subspace beamforming (AESB) is a commonly used method to set the dimensionality number in which signal subspace dimension is not equal to the number of sound sources but decided by a threshold adaptively because of the sparsity of speech signals in the time–frequency domain and the existence of noise.

(b) Diagonal loading

Diagonal loading adds a constant to the diagonal elements of \( R \). Its purposes are (1) to solve the irreversibility problem of \( R \); (2) to increase the robustness of the beamformer; and (3) to reduce the influence of eigenvectors with small eigenvalues.

Firstly, the white noise gain (WNG) is introduced as

\[ \text{WNG} = \frac{|w^H d|^2}{w^H w} \] (11)

Because when the noise is the spatial white noise or the channels are uncorrelated, \( R \) is a unit matrix, the denominator of Equation (11) represents the output power of noise. The research shows that the stability of adaptive beamforming can be effectively improved by controlling the signal-to-noise ratio (SNR) of the white noise, which is given as

\[ \min_w w^H R w, \text{WNG} = \delta^2, \ w^H d = 1 \] (12)

where \( \delta \) denotes the constraining value to be chosen that must be less than or equal to the maximum possible WNG, which is the number of microphones.

The solution of Equation (12) is

\[ w = \frac{[R + \varepsilon I]^{-1} d}{d^H [R + \varepsilon I]^{-1} d} \] (13)

where \( \varepsilon \) denotes the loading factor to control the angular loading, and the larger \( \varepsilon \), the larger the SNR of the white noise. The relation between \( \varepsilon \) and \( \delta \) is not simple, and it depends on the structure of the microphone array and frequency.

For MVDR, its adaptive enhancement performance is better with a smaller \( \varepsilon \), but its stability worsens. When \( \varepsilon \) is infinite, MVDR degenerates into DSB. For LCMV, the filter coefficients can be obtained as follows:

\[ w = \frac{[R + \varepsilon I]^{-1} C}{C^H [R + \varepsilon I]^{-1} C} g \] (14)

When \( \varepsilon \) is infinite, the direct inverse solution of the reverberation-free mixing process is obtained from Equation (14), and the adaptive beamforming degenerates to the
fixed beamforming. In addition to fixed diagonal loading, an expected SNR of the white noise can also be obtained by unidirectional searching of \( \varepsilon \); thus, the balance between performance and stability could be adjusted [8]. The authors in Reference [9] pointed out that the techniques of diagonal loading can also solve the mismatch problem of the steering vector with negative loading.

2.3 Independent component analysis

Independent component analysis (ICA) is one of the most widely used methods in separation algorithm [10]. Some studies show the adaptive beamformers are theoretically equivalent to ICA in the frequency domain [11]. The ICA decomposes and searches the signals using the properties of super-Gaussian and mutual independence of the sound sources [12, 13], and some new ICA algorithms can solve the ambiguity of permutation [14–16].

Firstly, the separated output signal in the time domain can be denoted as Equation (3), and then the equation in the frequency domain is

\[
Y(f) = W(f)X(f)
\]

where \( Y \) denotes the output signal; \( W \) is the separation matrix; \( X \) denotes the input signal. Among them, the initial value of \( W \) does not consider the interference factors of echoes and reverberations, as well as amplitude difference between two channels; therefore, it can be calculated with a spatial geometry method.

Suppose \( w \), a vector of \( W \), is the unmixing vector of sound source \( k \), \( yk = \omega^Hx \), where \( x \) is a vector of \( X \), is the separated signal of source \( k \). Therefore, if the separation matrix \( W \) of ICA is obtained, the separated signals of different sources can be considered as the enhanced signals of these sources, and the enhancement performance is determined by the evaluation precision of the separation matrix \( W \).

Moreover, ICA requires the number of separated signals to be equal to that of channels. Hence, it is usually applied along with PCA for dimensionality reduction when there are fewer separated signals than channels and for frequency bins with low SNRs.

Firstly, the cross-correlation matrix at each frequency bin is calculated by

\[
R_{XX}(f) = E[X(f)X(f)^H]
\]

The eigenvectors and eigenvalues corresponding to each frequency bin are obtained by decomposition of the cross-correlation matrix in the complex domain. If the number of active sound sources is \( N \), the eigenvalues are ordered in descending order, \( d_1 \geq d_2 \geq \cdots \geq d_N \) and the \( N \) eigenvalues above threshold and limited to the number of sound sources are selected to form a diagonal matrix \( D = \text{diag}(d_1, \ldots, d_N) \). The eigenvectors are combined into matrix \( E = [e_1, \ldots, e_N] \), and then the dimensionality matrix is given by

\[
V = D^{-1/2}E^H
\]

for which

\[
Z(f) = V(f)X(f)
\]

is the dimensionality-reduced signal. Matrix \( V \) whitens the signal, for example \( R_{ZZ} = I \), and hence simplifies iterations during ICA. The dimensionality of signal \( Z(f) \) is \( N \), which is the number of sound sources at the current frequency \( f \).

Then, the signal at each channel after dimensionality reduction can be expressed as

\[
Y(f) = W(f)V(f)X(f)
\]

Given the initial separation matrix, the ICA iterative process can be performed. Generally, the non-linear decorrelation is used, and the ICA recursion is expressed as [17, 18],

\[
W_{i+1}(f) = \eta[I - \langle \phi(y(f))y^H(f) \rangle_r] W_i(f) + W_i(f)
\]

where \( \eta \) is the iteration ranging from 0.1 to \( 10^{-3} \); \( i \) is the iteration index; \( y(f) = W_i(f) \times Z(f) \) is the separation result obtained from the \( i \)-th iteration matrix, \( \langle \cdot \rangle_r \) denotes the time averaging, and \( \phi(\cdot) \) denotes a non-linear function defined as

\[
\phi(y(f)) = \tanh(\text{Re}(y(f))) + j \cdot \tanh(\text{Im}(y(f)))
\]

where \( \text{Re}(\cdot) \) and \( \text{Im}(\cdot) \) denote the real and imaginary parts of their arguments, respectively.

The non-linear correlation is expressed as

\[
cor = \text{det}(I - \langle \phi(y(f))y^H(f) \rangle_r)
\]

where \( \text{det}(\cdot) \) denotes the determinant.

When the non-linear correlation of the signal is below a threshold or the iteration step exceeds the limit, ICA terminates.

3 PERFORMANCE EVALUATION INDICATORS OF SPEECH ENHANCEMENT

Quantitative indicators commonly used to evaluate speech enhancement performance include the SAR, the SDR, and the SIR [19, 20]. Some studies have shown that these indicators can quantify the performance of speech enhancement from various aspects. They also possess a common feature in that the larger these indicators are, the better the enhancement performance is.
The SAR is the ratio of the signal component and the noise introduced into the sound space. This indicator can describe the stability of a speech-enhancement method and the degree of loss of the target signal during processing. Its expression is given as

$$\text{SAR} = 10 \log_{10} \frac{\|s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}\|^2}{\|e_{\text{artif}}\|^2}$$ (23)

where $s_{\text{target}}$ represents the target speech component; $e_{\text{interf}}$ represents the interference component; $e_{\text{noise}}$ represents the noise component; and $e_{\text{artif}}$ represents the artefact error. Here, these components are obtained by projecting the processing results to the original signal because the projection method can tolerate the distortions of amplitude and the convolution with a fixed-length signal, and they do not require knowledge of the specific processing method.

SDR represents the ratio between the target component and the other components. The indicator is commonly used to evaluate the comprehensive performance of a speech-enhancement method. Its expression is given by

$$\text{SDR} = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{artif}} + e_{\text{interf}} + e_{\text{noise}}\|^2}$$ (24)

SIR represents the ratio between the target and the interference components. This indicator represents the performance of interference suppression. It can be formulated as

$$\text{SIR} = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}$$ (25)

In brief, these indicators describe the performance of the processed output signal relative to the original signal from different aspects. When two independent signals interfere with each other in the experiments, their power levels are adjusted to be the same. The SIR of the input signal is then equal to 0 dB, and the SIR of the output signal represents the SIR enhancement relative to the output signal. Therefore, the SIR of sound source $i$ can also be calculated as follows,

$$\text{SIR}_i = 10 \log_{10} \frac{\sum |x_i(t)|^2}{\sum |\sum_{k \neq i} x_k(t)|^2}$$ (26)

where $x_i$ and $x_k$ are the components of source $i$ and $j$, respectively.

The input SIR of the sound source $i$ at the receiving microphone $j$ is given by

$$\text{InputSIR}_i = 10 \log_{10} \frac{\sum |x_{ij}(t)|^2}{\sum |\sum_{k \neq i} x_{jk}(t)|^2}$$ (27)

where $y_{ij}(t)$ is the component of source $i$; $y_{jk}(t)$ is the residual components of other sources after enhancement.

The SIR of the enhanced signal is given by

$$\text{OutputSIR}_i = 10 \log_{10} \frac{\sum |y_{ij}(t)|^2}{\sum_j |\sum_{k \neq j} y_{jk}(t)|^2}$$ (28)

where $y_{ij}(t)$ is the component of source $i$; $y_{jk}(t)$ is the residual components of other sources after enhancement.

Finally, the SIR improvement of sound source $i$ can be expressed as

$$\text{ImproSIR}_i = \text{OutputSIR}_i - \text{InputSIR}_i$$ (29)

4 | EXPERIMENT AND ANALYSIS

4.1 | Experimental environment

To compare the performance of different speech-enhancement methods commonly used and establish a unified evaluation standard for speech enhancement, a group of speech-enhancement experiments were set up based on the Image Source Model (ISM) [21–23]. The experimental space spanned 6 m × 4 m × 3 m, the spatial coordinates of the microphone array were [3,0,2,0,1,5], and the microphone array was a 6 + 1 ring array in which one microphone was located in the centre and the other six microphones were evenly distributed in the circumference of a circle at equal separation distances equal to 4.35 cm. The red dots in Figure 1(a) denote the zero azimuth angles of the microphone array, the heights of the sound sources are equal to that of the microphone array, and the distance between the sources and the centre of the array microphone array is 1 m. Figure 1(b) is the structure of the microphone array.

In this study, we only evaluated the enhancement performance of an array with two sound sources. The position of one source was fixed to the zero-azimuth angle, and the other source rotated around the microphone array clockwise from 0 to 180°. Accordingly, a group of experimental data was collected when the second source rotated every 5°. Therefore, 37 groups of data were collected in total. The T60 parameter of our simulation environment was equal to 0 and 0.3 s, and the impulse response was calculated for various T60 parameters. There are 74 impulse response groups, and each group contained responses from seven channels. ISM still contained a slight reverberation to measure the impulse response when T60 was 0 s. In our experiment, we assumed that the microphone array only contained the direct waves transmitted by the sound sources. Setting T60 to 0.3 s emulates the acoustic scenes of most of the actual environments. Therefore, the performance evaluation of this condition is similar to real cases/applications. In this study, we do not add any noise to the signal mixing process, and there is no loss or distortion in the receiving process of the microphone array.
The data were obtained from the Texas Instruments Massachusetts Institute of Technology database. Five males and five females were selected. The lengths of all the signals are set to be at least 4 s to ensure the convergence of the speech-enhancement algorithms. The input powers of all the sound sources were the same; that is, the SNR was 0 dB, and no additional noise was added to the signals. All the algorithms were executed \(37 \times 2 \times 45 = 3330\) times, and we estimated the average enhancement value in the same environment as the performance evaluation outcome of each tested algorithm.

The speech-enhancement algorithms in this study included DSB, MVDR, LCMV, and ICA. Among them, DSB included the DSB version and the DSB with a binary mask version, while MVDR and LCMV included two implementations based on diagonal loading and feature space decomposition. In turn, the DSB and MVDR algorithms supported online implementations. For the online MVDR, the objective was to minimize the output power of the current frame, and the WNG was the iteration boundary according to reference [8]. Additionally, the alignment centre of the time axis of the DSB was the array centre, and the processing that was based on the binary mask only retained the time–frequency bins with the maximum power, while the other time–frequency bins were set to zero. The basic MVDR used diagonal loading with minimum diagonal loading elements. In the derivation method, the suffix with ‘AESB’ was the eigendecomposition, while the upper limit of the dimensionality reduction was the number of sound sources. Additionally, the suffix ‘search’ denoted the diagonal loading in which WNG was maintained iteratively within a specified range, and the spatial pointing characteristic of the filter was maximized. Moreover, LCMVs had a null constraint on the interference source except for the non-distortion constraint of the target. The basic LCMV used diagonal loading with minimal diagonal loading elements. The derivative of LCMV was similar to MVDR and contained fixed beamforming with maximum diagonal loading elements. In addition, the ICA algorithm in this study was PCA + ICA, and it used DSB and PCA to initialize the de-mixing matrix so that the uncertainty of the convergence order of ICA could be avoided.

The entire experimental protocol is listed in Table 1.

## 4.2 Experimental results

### 4.2.1 Experiment I. Performance comparison of the basic enhancement algorithms

(a) \(T60 = 0\) s

DSB, MVDR, LCMV, and ICA, were tested in this experiment. The MVDR and LCMV were combined with diagonal loading, and the loading factor was the minimal diagonal element. The experimental results are shown in Figure 2, whereby Figure 2 (a)–(c) depicts the SDR, SIR, and SAR after enhancement. In

---

**FIGURE 1** Schematic diagram of our simulation environment and the microphone array

| Data Source | TIMIT |
|-------------|-------|
| Data type, signal length | Five males and five females, at least 4 s |
| Microphone array | 6 + 1 ring array |
| Sound source number | 2 |
| Angle variation between two sound sources | \(0\)°–\(180\)°, step \(5\)° |
| Algorithms | DSB, MVDR, LCMV, and ICA |
| Experimental environment | \(T60 = 0\) s and \(T60 = 0.3\) s |
| Execution times | \(37 \times 2 \times 45 = 3330\) |

Abbreviations: DSB, delay-and-sum beamforming; ICA, independent component analysis; LCMV, linearly constrained minimum variance; MVDR, minimum variance distortionless response; TIMIT, Texas Instruments Massachusetts Institute of Technology.
addition, the horizontal axis is the angle between two sound sources, the vertical axis is the enhancement indicator value, and the units are in dB.

From Figure 2, it can be observed that LCMV and ICA can effectively suppress the interference source when the azimuth of the interference source is known because their SDR and SIR indicators are higher than those of the other methods. Although the enhancement of MVDR and DSB increases as a function of the angle between the two sources, it is still unable to obtain higher noise suppression because the objective of MVDR is to minimize the output power, and its enhancement performance is limited by the following factors: (1) large diagonal loading limits the degree of freedom of MVDR owing to the requirement of uncorrelated noise suppression and (2) the summation output power of the two speech sources may be smaller than the powers of any of the two individual sources because of power cancelation.

In addition, the performance of the ICA and LCMV decreases when the interval angle is 180° because the ISM model contains reflection wave components at T60 = 0 s when two sound sources are symmetrical relative to the microphone array. It is impossible to eliminate the reflection wave components even if their respective directions are known.

(b) T60 = 0.3 s
This experiment compared the DSB, MVDR, LCMV, and ICA methods, for T60 = 0.3 s. MVDR and LCMV were improved by diagonal loading using the minimum diagonal element. The experimental results are shown in Figure 3, whereby Figure 3 (a)–(c) depict the SDR, SIR, and SAR after enhancement, the horizontal axis is the interval angle between two sound sources, the vertical axis is the enhanced indicator value, and the units are in dB.

From Figure 3, we can observe that the SDR and SAR do not have high values because the reverberation components have been introduced into the noise that was accounted for in the calculation of these enhancement methods. Therefore, we evaluated the enhancement performance of the various methods according to the SIR. Firstly, compared with DSB and MVDR, LCMV and ICA are more effective in suppressing the interference sources, reaching values that are approximately equal to 20 dB. However, the ICA improves the performance gradually in the range of 0–90°, but its performance is worse than that of LCMV. The reason is attributed to the fact that in reverberation, the steering vectors of the sound sources with similar angle intervals are close to each other, and it is very difficult to effectively control the final convergence order of
the algorithms by presetting the initial points of iteration. This ultimately results in a decline in ICA performance when the difference of the angles is small.

Secondly, the SIR values of the MVDR and DSB do not change much relative to the case of $T60 = 0\,\text{s}$ because they are able to suppress diffused and incoherent noise, and the reverberation components can be considered diffused noise.

4.2.2 | Experiment II. Performance comparison of adaptive beamforming methods

(a) $T60 = 0\,\text{s}$

This experiment tested the enhancement performances of MVDR and LCMV that improved following the applications of eigendecomposition and diagonal loading when $T60 = 0\,\text{s}$. The suffixes of MVDR included four types: ‘AESB’, ‘Search’, ‘PCA’, and ‘offline’, where ‘AESB’ denotes the adaptive dimensionality selection, ‘Search’ denotes the WNG target iteration value with diagonal loading (and the unmarked WNG value is 0 dB), ‘PCA’ denotes the dimensionality reduction with a fixed dimension of two, and ‘offline’ denotes the basic MVDR. The experimental results are shown in Figure 4, where Figure 4(a)–(c) is the enhanced SDR, SIR, and SAR, respectively, and the horizontal axis is the interval angle between the two sound sources. The longitudinal axis is the enhancement indicator value in units of $\text{dB}$.

The following conclusions can be drawn from Figure 4:

1. When the fixed dimensionality number of PCA is equal to the number of the sound sources, none of the enhancement indicators are satisfactory because the number of effective components in some frequency bins is less than two owing to the sparsity of speech signals. If we do not increase the diagonal loading and directly solve Equation (10) in the two-dimensional space, the irreversibility of the covariance matrix leads to strong distortion and enhancement outcomes. AESB uses an adaptive dimensional reduction to make the algorithm solvable at all the frequency bins in the reversible space that are composed of the effective components. Accordingly, the result is satisfactory even with the use of a small interval angle. It

FIGURE 3  Enhancement effect of the basic enhancement methods when $T60 = 0.3\,\text{s}$. DSB, delay-and-sum beamforming; ICA, independent component analysis; LCMV, linearly constrained minimum variance; MVDR, minimum variance distortionless response.
FIGURE 4  Enhancement effect of the adaptive enhancement methods when $T_{60} = 0$ s. AESB, adaptive eigen-subspace beamforming; LCMV, linearly constrained minimum variance; MVDR, minimum variance distortionless response; PCA, principal component analysis, Search, diagonal loading when the white noise gain equals to 0 dB; Search -5, diagonal loading when the white noise gain equals to -5 dB; offline, the basic MVDR.

can be concluded that the iteration goal of MVDR is not a problem. However, dimensionality has a considerable impact on the enhancement outcomes.

(2) The enhanced performance of the basic MVDR decreases when the angle difference between the sources is small. The reason is attributed to the fact that its degree of freedom falls into the non-signal space because of the large diagonal loading factor. However, the performance of MVDR can be improved by adjusting diagonal loading. Correspondingly, the smaller the diagonal loading factor, the smaller the WNG. This means that the suppression of the interference sources increases, and the stability decreases. In theory, the factors influencing the diagonal loading technology include the following. (1) Angle differences between sound sources: When the WNG value is fixed, the angle difference is inversely proportional to the loading value; that is, the enhancement performance is different at different angle intervals. (2) Frequency: Although the target WNG value of diagonal loading is the same at each frequency bin, the overlap of the signals is inconsistent owing to the sparsity of the speech signals in the time–frequency domain, which results in a different diagonal loading requirement to obtain the same WNG. Therefore, this affects the enhancement of the performance. In summary, the relation between the loading and WNG values is related to frequency and angle, making it difficult to choose. Thus, if the stability is considered, the suppression performance is weakened. Figure 4 shows that in the instance when the value of WNG is equal to −5 dB, the inhibition effect is better than that evoked for a value of 0 dB.

(3) LCMV elicits the best performance when it is loaded diagonally. When there is no echo and the accurate steering vectors of each sound source are known, the constraint space can guarantee the suppression performance, and the diagonal loading only needs to ensure the feasibility of the enhancement methods. In this case, both the eigendecomposition and the diagonal loading techniques can improve the enhancement performance. However, because the eigendecomposition still needs to consider the dimensionality reduction, it will be affected by the accuracy of the adaptive dimensionality reduction when the angle
difference is small. Accordingly, the enhancement of the performance will be slightly decreased.

(b) $T_{60} = 0.3$ s

This experiment tested the improved MVDR and LCMV methods based on eigendecomposition and diagonal loading when $T_{60} = 0.3$ s. The experimental results are shown in Figure 5.

The following conclusions can be drawn from Figure 5:

1. In a reverberation environment, the enhancement performance of LCMV and MVDR is limited by the eigendecomposition, especially when the sources are approaching symmetrical positions. This is because the strong reverberation interference leads to the inability of eigen decomposition to determine the appropriate dimensionality reduction number. The performance of diagonal loading in the reverberation environment shows improved stability, the SIR of LCMV is still high owing to the accurate suppression of the direct wave components of the interference sources, and MVDR also improves as the diagonal loading factor increases. When the angle difference is greater than 90°, the diagonal loading with a WNG value of $-5$ dB makes MVDR obtain a similar SIR compared with LCMV. When the angle difference is low, the SIR of the MVDR decreases, and the SAR is more stable because LCMV does not restrict the azimuth angle of the interference source. Therefore, MVDR exhibits an improved anti-reverberation capacity compared with LCMV.

2. According to the comprehensive evaluation of SDR, the MVDR with a WNG value of $-5$ dB performs best in the reverberation environment.

4.2.3 Experiment III. Performance comparison of online algorithms

Herein, ‘online’ indicates that the enhancement algorithms iterate once for a single frame of the speech signal. Therefore, compared with offline processing, the advantage of online processing relies on the fact that the iteration data constitutes...
the current frame. In this sense, processing can make better use of the short-term stability of the speech signals; that is, it is not affected by signal mutations. Therefore, this is more meaningful for online performance testing when improved real-time performance is required.

(a) T60 = 0 s
This experiment tested the online enhancement performances of DSB + MASK and MVDR when T60 = 0 s. To compare these, we also calculated the offline indicators of MVDR and LCMV. The experimental results are shown in Figure 6. As shown in this figure,

(1) DSB + MASK yields the highest SIR value and possesses a strong capacity to suppress interference sources but also leads to the greatest signal loss. Therefore, the performance of SDR is not outstanding. However, because the angle information of the interference source is used, the performance of each angle is relatively stable.

(b) T60 = 0.3 s
This experiment tested the online enhancement performance of DSB + MASK and MVDR when T60 = 0.3 s and the

(2) When the WNG boundary is −5 dB, the interference suppression ability of the online MVDR is better than 0 dB, but the signal distortion follows an opposite trend, that is, 0 dB is better than −5 dB. Compared with the offline iterative MVDR, the performance of the online MVDR is slightly worse. The reason is attributed to the fact that the online algorithm only iterates once for each frame to ensure the minimum amount of computation. This makes it difficult to achieve a stable point of iteration. In general, the offline versions of speech enhancement can be used to evaluate the relative limitations of these algorithms, while the online versions reflect their real-time tracking capacities that are affected by the iteration method and the iteration step size.

**Figure 6** Enhancement effect of online algorithms at T60 = 0 s. DSB, delay-and-sum beamforming; DSB-Mask, delay-and-sum beamforming with a binary mask; LCMV, linearly constrained minimum variance; MVDR, minimum variance distortionless response; Search, diagonal loading when the white noise gain equals to 0 dB; Search -5, diagonal loading when the white noise gain equals to -5 dB; online0, the enhancement algorithms iterate once for a single frame of the speech signal when the white noise gain equals to 0 dB; online-5, the enhancement algorithms iterate once for a single frame of the speech signal when the white noise gain equals to -5 dB.
offline indicators of MVDR and LCMV. The experimental results are shown in Figure 7. As shown in this figure,

1. In the reverberation environment, the SIR of DSB + MASK is higher than that for LCMV. Given that the MASK-based methods are non-linear, it is difficult to compare them with the linear methods in theoretical analyses. However, it can be found that the hard classification method based on the maximum power has unique features in interference removal. Additionally, this method has similarities with the spectral subtraction methods in the single-channel case, but their theoretical bases are different.

2. The mask methods based on the output power of DSB are superior to the methods based on power classification. Correspondingly, the MASK-based methods also cause considerable signal losses. Accordingly, the SAR is relatively low, and the comprehensive indicator of SDR is not prominent.

3. The online version of MVDR is more stable than the offline version in the reverberation environment, and it does not decrease significantly. Moreover, both the online and offline versions have a specific anti-reverberation capacity, so the SAR is relatively high. As for SDR, the online version of MVDR and DSB + MASK have similar performances.

4.2.4 Experiment IV: Comprehensive performance

Comprehensive performance refers to the selection of methods with a better enhancement effect as shown in the experimental Sections 1–3 to compare their various performance indicators under different reverberation conditions.

(a) T60 = 0 s
This experiment tested the enhancement effects of DSB + MASK, ICA, the online MVDR, and MVDR and LCMV with diagonal loading. The experimental results are shown in Figure 8.
As shown, LCMV and ICA yield the best performances in the non-reverberation environment. Specifically, the SIR of ICA is very high. However, its SAR is lower than that of LCMV. The reason is that LCMV is a method based on direct solving, while ICA is based on numerical iteration. Therefore, ICA can achieve a better suppression effect owing to its gradual approach to the signal distribution model, but its signal loss is also high. In general, LCMV and ICA have similar performances in reverberation-free environments.

Although the combined DSB + MASK method yields higher SIR values than LCMV, its comprehensive performance is limited because of the tremendous signal loss. Given that MVDR only uses the information of the target sound source, its performance decreases at small angle intervals, and the result of the offline version is better than the online version. Furthermore, if there are steering vector mismatches, a signal cancelation of the online MVDR will occur significantly with the occurrence of signals, and the entire signal power will gradually decrease. This phenomenon does not occur in the case of offline algorithms. Therefore, the capacity of the online version to achieve signal suppression is increased.

(b) T60 = 0.3 s
This experiment tested the enhancement effect of DSB + MASK, ICA, the online MVDR, the MVDR, and the LCMV with diagonal loading when T60 = 0.3 s. The experimental results are shown in Figure 9. The following conclusions can be drawn from Figure 9:

(i) In the reverberation environment, the SIR value of DSB + MASK is the highest. If the order confusion problem with a low-angle difference can be solved, the interference suppression ability of ICA is similar to that of LCMV. Except for the problem of confusion and power loss of ICA, long-tail reverberation can also make it inapplicable to the mixing process of signal sources because the delayed signal is then mixed into the new signal, and the frame length of the Fourier transform is difficult to cover the entire reverberation process.
Therefore, if we can improve the modelling of the hybrid process, the performance of ICA should be significantly improved in the reverberation environment.

(2) LCMV and MVDR restrict the signal subspace and search for the minimum output power in the non-signal subspace. This process requires these methods to simultaneously suppress reverberation, diffused noise, and uncorrelated noise by eliminating the output power. Accordingly, the output power is set as the iteration. Therefore, the final processing outcome is the comprehensive suppression of all these interfering components. The problem associated with lack of pertinence causes disadvantages to the spatial search. If the reverberation process can be included in the acoustic transfer function estimation or space searches without steering vectors, the effect should be improved.

Finally, a statistical analysis of all the experiments was conducted, and the result was as follows.

(1) In the performance comparison of the basic enhancement algorithms including DSB, MVDR, LCMV, and ICA, LCMV and ICA can effectively suppress the interference source with their known azimuths.

(2) In the performance comparison of adaptive beamforming methods including MVDR, LCMV, and the improvement methods based on them, LCMV elicits the best performance with diagonal loading when T60 = 0 s, while MVDR exhibits an improved anti-reverberation capacity when T60 = 0.3 s.

(3) In the performance comparison of the online algorithms including DSB + MASK and MVDR when T60 = 0.3 s, as well as the offline indicators of MVDR and LCMV, DSB + MASK yields the highest SIR value and possesses a strong capacity to suppress interference sources, and the online version of MVDR is more stable than the offline version in the reverberation environment.

(4) As for the comprehensive performance of DSB + MASK, ICA, the online MVDR, and MVDR and LCMV with
ICA yielded the best effects when there was no reverberation. The comprehensive performance test showed that LCMV and SIR and possessed the capacity to suppress interference [24]. DSB showed better performance. For online enhancement, the reverberation. However, if the hybrid modelling process of the improved MVDR and LCMV, were studied in the indoor environment of $T60 = 0$ s and $T60 = 0.3$ s. The results showed that the enhancement effects of LCMV and ICA were relatively more stable as basic enhancement methods irrespective of reverberation. In the case of the improved enhancement algorithms, methods that employed diagonal loading iterations showed better performance. For online enhancement, the DSB + MASK method yielded the best performance on the SIR and possessed the capacity to suppress interference [24]. The comprehensive performance test showed that LCMV and ICA yielded the best effects when there was no reverberation, while DSB + MASK yielded the best SIR value in the case of reverberation. However, if the hybrid modelling process of the ICA method or the space search method of LCMV and MVDR can be improved, these methods will be amenable to further refinement and use.

### 5 | CONCLUSIONS

This study tested the enhanced performance of basic enhancement methods, including DSB, MVDR, LCMV, and ICA. The adaptive enhancement methods, including the improved MVDR and LCMV, were realized by eigendecomposition and diagonal loading. In addition, the performance of online enhancement based on the iteration of single-frame speech signals, and the comprehensive performance of various enhancement methods, were studied in the indoor environment of $T60 = 0$ s and $T60 = 0.3$ s. The results showed that the enhancement effects of LCMV and ICA were relatively more stable as basic enhancement methods irrespective of reverberation. In the case of the improved enhancement algorithms, methods that employed diagonal loading iterations showed better performance. For online enhancement, the DSB + MASK method yielded the best performance on the SIR and possessed the capacity to suppress interference [24]. The comprehensive performance test showed that LCMV and ICA yielded the best effects when there was no reverberation, while DSB + MASK yielded the best SIR value in the case of reverberation. However, if the hybrid modelling process of the ICA method or the space search method of LCMV and MVDR can be improved, these methods will be amenable to further refinement and use.

### ACKNOWLEDGEMENTS

This work was supported by the Natural Science Foundation of China under Grant 61973059.

### ORCID

Yangjie Wei □ https://orcid.org/0000-0001-6615-5484

### REFERENCES

1. Wang, D.L., Chen, J.T.: Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. 26(10), 1 (2018)
2. Hiroshi, S.: Multichannel extensions of non-negative matrix factorization with complex-valued data. IEEE Trans. Audio Speech Lang. Process. 21(5), 971–982 (2013)
3. Kitamura, D., et al.: Determined blind source separation with independent low-rank matrix analysis. In: Makino S. (ed.), Audio Source Separation. Signals and Communication Technology, pp. 125–155. Springer, Cham (2018)
4. Takaya, H.: Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR. IEEE/ACM Trans. Audio Speech Lang. Process., 25(4), 780–793 (2017)
5. Warsitz, E., Haeb-Umbach, M.R.: Blind acoustic beamforming based on generalized eigenvalue decomposition. IEEE Trans. Audio Speech Lang. Process. 15(3), 1529–1539 (2007)
6. Anastasios, A., Griffin, A., Mouchtaris, A.: Capturing and reproducing spatial audio based on a circular microphone array. J. Electr. Comput. Eng. 2003, 1–16 (2013)
7. Feldman, D.D., Griffiths, L.J.: A projection approach for robust adaptive beamforming. IEEE Trans Signal Process 42(4), 867–876 (1994)
8. Cox, H., Zeskind, R., Owen, M.: Robust adaptive beamforming. IEEE Trans. Acoust. Speech Signal Process. 35(10), 1365–1376 (1987)
9. Vincent, F., Besson, O.: Steering vector errors and diagonal loading. Radar Sonar Navig. IEEE Proc. 151(6), 337–343 (2004)
10. Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic, San Francisco (2010)
11. Shoko, A.: Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures. EURASIP J. Adv. Signal Process. 2003(31), 1157–1166 (2003)
12. Sawadah, H.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech Audio Process. 12(5), 530–538 (2004)
13. Ikeda, S., Murata, N.: A method of ICA in time-frequency domain. In Proceedings International Workshop on ICA & BSS, pp. 365–371. Aussois (1999)
14. Atsuo, H.: Solution of permutation problem in frequency domain ICA, using multivariate probability density functions. In: International Conference on Independent Component Analysis & Blind Signal Separation, pp. 601–608. Springer-Verlag, Charleston (2006)
15. Taets, K.: Blind source separation exploiting higher-order frequency dependencies. IEEE Trans. Audio Speech Lang. Process. 15(1), 70–79 (2007)
16. Nobuoka, T.: Stable and fast update rules for independent vector analysis based on auxiliary function technique. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2011, IEEE, New Paltr, 16–19 (Oct 2011)
17. Anthony, B.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7(6), 1129–1139 (1995)
18. Amari, S.: Natural gradient works efficiently in learning. Neural Comput. 10(2), 251–276 (1998)
19. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. 14(4), 1462–1469 (2006)
20. Févotte, C., Gribonval, R., Vincent, E.: BSS eVAL toolbox user guide. IRISA, Tech. Rep. 1706 (2005)
21. Lehmann, E.A., Johansson, A.M.: Diffuse reverberation model for efficient image-source simulation of room impulse responses. IEEE Trans. Audio Speech Lang. Process. 18(6), 1429–1439 (2010)
22. Ward, B.D., Lehmann, E.A., Williamson, R.C.: Particle filtering algorithms for tracking an acoustic source in a reverberant environment, IEEE Trans. Speech Audio Processing. 11(6), 826–836 (2003)
23. Gal, R., Gannot, S., Cohen, L.: Performance analysis of dual source transfer-function generalized sidelobe canceller. Speech Commun. 49(7–8), 602–622 (2007)
24. Wei, Y.J., et al.: Comparison of conventional enhancement and separation methods in indoor environments. In: Proceedings of 9th International Conference on Computing and Pattern Recognition (ICCPR), pp. 383–389. Xiamen (2020)

How to cite this article: Wei Y, Zhang K, Wu D, Hu Z. Exploring conventional enhancement and separation methods for multi-speech enhancement in indoor environments. Cogn. Comput. Syst. 2021;1–16. https://doi.org/10.1049/ecs2.12023