IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 28, 2021, accepted September 8, 2021, date of publication September 16, 2021,
date of current version September 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3113457

Unique Methods for Determining the Attenuation and Delay in Blind Source Separation Based on the Degenerate Unmixing Estimation Technique

KUANG-YOW LIAN, (Member, IEEE), AND JIA-HSIN LIN
Department of Electrical Engineering, National Taipei University of Technology, Taipei 10608, Taiwan
Corresponding author: Kuang-Yow Lian (kylian@mail.ntut.edu.tw)

This work was supported by the Ministry of Science and Technology (MOST), Taiwan, under Grant 109-2221-E-027 -110.

ABSTRACT The degenerate unmixing estimation technique (DUET), enables facile sound source separation through two signal mixtures. However, in numerous practical situations, identifying peaks in DUET histograms is difficult, which in turn prevents effective source separation. Specifically, sound sources with short time percentages in mixtures are more unlikely to form peaks in a histogram. Besides, when the noise floor of mixtures is high, the height of the histogram peak may be reduced, which causes peak position bias and prevents effective source separation. In this study, unique methods are proposed to solve the aforementioned problems. A subhistogram is established for each time frame; thus, the original histogram is divided into a series of subhistograms. Two measurement indices are then proposed to identify the accurate attenuation and delay parameters for sound sources: maximum distribution, which is based on the infinity norm, and variance distribution, which is based on variances. The results verify that maximum distribution effectively highlights the peaks of instantaneous sound sources when the time percentage difference between sound sources is excessively high. In this scenario, the conventional histogram is unable to find the peaks. When a strong noise disturbance is present in mixtures, variance distribution can be employed to estimate peak positions with lower biases than those in the conventional histogram. In addition to more well-defined and intense peaks, the proposed method can also reduce the bias error by 71% when the SNR is 0 dB. Variance distribution is more robust against noise disturbance compared to conventional histograms and can be adopted in a wide range of applications.

INDEX TERMS Attenuation and delay histogram, blind source separation, DUET, instantaneous sound sources, W-disjoint orthogonality.

I. INTRODUCTION

Blind source separation (BSS) refers to the separation of unknown source signals from mixtures of signals [1]–[3]. BSS is typically performed through independent component analysis (ICA) [4]–[7]. In addition to ICA, many other BSS methods convert mixtures to the time-frequency (TF) domain and then use TF masking to separate sound sources [8], [9]. Effective sound source separation is typically achieved using the degenerate unmixing estimation technique (DUET), which involves analyzing the attenuation and delay between two mixtures to reconstruct and separate all unknown sound sources [10]–[12]. Moreover, the environmental sound recorded using two microphones is similar to the auditory scenes perceived through human ears. Therefore, the DUET has been adopted in a wide range of applications, including automation field [13]–[15] and smart living [16], [17].

Ideally, the DUET requires the W-disjoint orthogonality (W-DO) of sound sources in the TF domain, meaning that all the sound must belong to only one source at any TF point. Consequently, the sound sources have a considerably

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

sparse distribution and exhibit nearly no overlaps at any TF point. Thus, distortion can be prevented when sound sources are separated through TF masking. Most existing TF masking methods for sound source separation are based on this hypothesis. Rather than equality between the number of sources and the number of mixtures, DUET-based methods only require source sparsity. When this condition is fulfilled, the problem in which the number of sound sources is higher than that of mixtures can be solved. Accordingly, the DUET is a viable solution for the aforementioned problem.

The DUET involves collecting attenuation–delay pairing data at all TF points and converting the attenuation to symmetric attenuation to establish a two-dimensional (2D) symmetric attenuation–delay histogram. All the peak positions in the histogram are then estimated to identify the attenuation–delay pairing parameters for each unknown sound source. These parameters enable the determination of which sound source each TF-point belongs to through maximum likelihood analysis and the reconstruction of sound sources through TF. Creating a satisfactory histogram is a crucial step in the DUET. Comprehensive sound source separation requires the accurate determination of the peak values and peak positions on a histogram. In fact, there are many situations that may lead to the failure of establishing histograms. In [21], the influence and solutions caused by phase wrap were mentioned, but there are still some situations that have not been explored by literature so far. For instance, the higher the activity time percentage of a sound source, the higher is the sound's peak value in a histogram. If another sound source with a low activity time percentage exists at this time, its peak value will be very low and difficult to detect using unsupervised peak searching methods. In addition, when mixtures contain excessive noise, the peak value tends to deviate from its correct position, decrease, or be submerged by the surrounding sound. Although numerous methods have been developed to reduce noise, reducing noise in mixtures is impossible, which renders noise disturbance a tricky problem.

Ideally, when no sound sources exist within a short period, the bar of any bin in the histogram has a height of 0. The heights of the bars change only when a sound source appears. However, a DUET histogram is generated through the accumulation of full-time data, and the details on the changes in each time frame are not recorded or analyzed. To observe the changes in the DUET histogram in short periods, a subhistogram was established in this study for each time window frame through short-time Fourier transform. Two metric indices, namely maximum distribution and variance distribution were devised according to the set of subhistograms to identify the correct number of sound sources and peak positions. Maximum distribution refers to the maximal instantaneous height in a subhistogram, which is obtained using the infinity norm to eliminate problems related to the insufficient peak values of an instantaneous sound source compared with those of other sources in the original histogram. Variance distribution refers to the differences between subhistograms according to variance calculations. It enhances the accuracy of peak position estimation and the robustness of mixtures against noise disturbance. Maximum distribution and variance distribution are applicable to general circumstances in addition to the aforementioned problems.

II. PRELIMINARY INFORMATION ON THE DUET

This section describes the sound source separation processes in the DUET and the problems encountered in these processes.

A. W-DO OF SOURCES

To conduct source estimation with the DUET, the W-DO criterium must be satisfied as much as possible for the sound sources under the assumption that any TF point in mixtures belongs to only one source. Specifically, the product of any two sound sources at any TF point equals 0, as shown below:

$$\hat{S}_f(\tau,\omega)\,\hat{S}_g(\tau,\omega)=0,\quad \forall(\tau,\omega),\ \forall f\neq g \tag{1}$$

where $\hat{S}_f(\tau,\omega)$ and $\hat{S}_g(\tau,\omega)$ refer to the values of sound sources f and g at the TF point (τ,ω), respectively. If W-DO is completely satisfied, nearly no distortion occurs in source separation based on TF masking. Therefore, greater sound source sparsity leads to more complete satisfaction of W-DO.

B. PARAMETERS OF THE ATTENUATION AND DELAY MODEL

In the DUET, only the attenuation and phase delay caused by the direct path from sound sources to two mixtures are considered, whereas the reverberation caused by the sources is ignored. The attenuation and phase difference for a sound source between two mixtures are expressed as follows:

$$\begin{aligned} x_1(t) &= \Sigma_{n=1}^{N} s_n(t) \\ x_2(t) &= \Sigma_{n=1}^{N} \bar{a}_n s_n\left(t-\bar{\delta}_n\right) \end{aligned} \tag{2}$$

where $(x_1,\ x_2)$, s_n, and N represent the mixtures, source signal, and number of sources, respectively, and $\bar{a}_n$ and $\bar{\delta}_n$ represent the attenuation and phase delay of the n^{th} source signal s in the second mixture, respectively. According to a narrowband assumption [19], Equation (2) can be expressed as follows to identify the relative relationship between two mixtures:

$$\begin{bmatrix} \hat{x}_1(\tau,\omega) \\ \hat{x}_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ \bar{a}_1 e^{-i\omega\bar{\delta}_1} & \cdots & \bar{a}_N e^{-i\omega\bar{\delta}_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(\tau,\omega) \\ \vdots \\ \hat{s}_N(\tau,\omega) \end{bmatrix} \tag{3}$$

When the TF point belongs to the n^{th} sound source, (3) means that the ratio of the sound source reflected in $\hat{x}_2$ and $\hat{x}_1$ will be $\bar{a}_n e^{-i\omega\bar{\delta}_n}$, where $\bar{a}_n$ and $\bar{\delta}_n$ are unknown constants. The preliminary steps involved in DUET are calculation of ratio of the measured values $\hat{x}_2$ to $\hat{x}_1$, estimation of the attenuation and

delay parameters on all TF points, and then find the positions where the parameters appear the most times. For more details, see [20].

On the basis of (3), the DUET attenuation and delay parameters for each TF point can be instantaneously estimated as follows:

$$a(\tau,\omega) := \left|\frac{\hat{x}_2(\tau,\omega)}{\hat{x}_1(\tau,\omega)}\right| \\ \delta(\tau,\omega) := \left(-\frac{1}{\omega}\right)\angle\frac{\hat{x}_2(\tau,\omega)}{\hat{x}_1(\tau,\omega)} \quad (4)$$

where $a(\tau,\omega)$ and $\delta(\tau,\omega)$ are the estimated value of attenuation and delay at each TF point. The most frequent (a,δ) values can be used as the estimated values of $\bar{a}_n$ and $\bar{\delta}_n$. The aforementioned steps are the basis of the DUET. When a single sound source is close to sensor x_1, the attenuation $a(\tau,\omega)$ ranges from 1 to 0. When this source is closer to x_2 than to x_1, the attenuation ranges from 1 to ∞. This is not conductive to the subsequent statistical inference. Consequently, $a(\tau,\omega)$ is substituted with symmetric attenuation by using the following equation [11]:

$$\alpha(\tau,\omega) = a(\tau,\omega) - 1/a(\tau,\omega) \quad (5)$$

where α indicates symmetric attenuation. After $a(\tau,\omega)$ is substituted with symmetric attenuation, the attenuation values ranging from 0 to 1 become negative and those ranging from 1 and ∞ become positive. The coordinate axis in terms of symmetrical attenuation can better present the distance of the sound source. On the basis of (5), the attenuation of a sound source $(\bar{a}_n)$ is defined as follows: $\bar{\alpha}_n$=$\bar{a}_n - 1/\bar{a}_n$.

C. ATTENUATION DELAY HISTOGRAM

The estimated $\alpha(\tau,\omega)$ and $\delta(\tau,\omega)$ values at each TF point are expressed as (α,δ) paired values. All the paired (α,δ) data are clustered to identify N centers. Theoretically, the number of clusters and the positions of their centers constitute the satisfactory estimation of the number of unknown sources and the corresponding attenuation and phase difference values $\left(\bar{\alpha}_n,\bar{\delta}_n\right)$. Notably, to prevent phase ambiguity, the following restriction condition must be fulfilled:

$$|\omega\delta| < \pi, \quad \forall\omega \quad (6)$$

When $\omega\delta$ exceeds π, phase wrap occurs, which causes a major bias in the delay value and diminishes its relevance. Therefore, to ensure that the delay value can be applied in a logical manner, the sampling rate and the distance between two sensors must be controlled in the sampling process. The higher the sampling rate, the shorter must be the distance between two sensors. Let the speed of sound in air be c = 344 m/s and the sampling frequency be $2\cdot\omega_{\max}$ ($\omega_{\max}$ is the maximal detectable spectral frequency and is half the sampling frequency). According to (6), the distance between two sensors must be shorter than $\pi c/\omega_{max}$. For example, if the sampling frequency is 8,000 Hz, the distance between two sensors must be shorter than 8.59 cm.

After all the paired (α,δ) values are calculated using (4) and (5), the number of unknown sound sources and the corresponding attenuation and phase differences must be determined. The most efficient method for performing the aforementioned task is to plot a 2D $\alpha\delta$ histogram, with the number of peaks representing the number of unknown sources and the peak positions representing the paired parameters of the sources. According to the principle of maximum likelihood, the weighted formula $\left|\hat{x}_1(\tau,\omega)\,\hat{x}_2(\tau,\omega)\right|^p\omega^q$ can be derived, where p and q are the weight parameters. In a few special occasions, we can adjust the weight parameters to make the sound source more prominent on the histogram. Subsequently, we merge the histogram with the weighted formula, a 2D weighted histogram can be expressed as follows [10], [11]:

$$H_{\Delta\alpha,\Delta\delta}(B,D) = \Sigma_{\tau,\omega}M_{B,\Delta\alpha}(\tau,\omega)\,M_{D,\Delta\delta}(\tau,\omega) \\ \times\left|\hat{x}_1(\tau,\omega)\,\hat{x}_2(\tau,\omega)\right|^p\omega^q \quad (7)$$

where $H_{\Delta\alpha,\Delta\delta}(B,D)$ represents the weighted histogram; B and D represent the bin indices of α and δ, respectively ($B = 1, 2, \ldots, k$; $D = 1, 2, \ldots, l$); p and q are weight parameters; and $\Delta\alpha$ and $\Delta\delta$ are the resolutions of B and D in the histogram, respectively. Further, $M_{B,\Delta\alpha}$ and $M_{D,\Delta\delta}$ refer to the α and δ selection conditions, respectively, and are expressed as follows:

$$M_{B,\Delta\alpha}(\tau,\omega) = \begin{cases}1 & \text{if } |\alpha(\tau,\omega) - B| < \Delta\alpha \\ 0 & \text{otherwise}\end{cases} \\ M_{D,\Delta\delta}(\tau,\omega) = \begin{cases}1 & \text{if } |\delta(\tau,\omega) - D| < \Delta\delta \\ 0 & \text{otherwise}\end{cases} \quad (8)$$

An adequate histogram is the key for accurately calculating α and δ. Fig. 1 illustrates the plotted histogram.

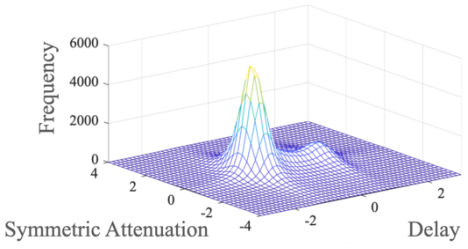


FIGURE 1. Weighted histogram established using real recording. Frequency labeled at z-axis is the number of occurrences. In real applications, B and D are calculated over wider ranges than those presented in this figure to increase the resolution (e.g., $B = 1, \ldots, 125$; $D = 1, \ldots, 125$).

Ideally, one should be able to detect a slender peak from an unknown source in the plotted histogram. The number of peaks in the histogram represents the number of unknown sound sources. The 2D coordinate of the peak represents the paired attenuation–delay parameters of an unknown source. After the paired parameters are identified, the sound sources are reconstructed through maximum likelihood analysis [10].

D. AMBIGUITY OF SEARCHING PEAKS

The contents of many of the mixtures at each TF point comprise values outside the direct path between the sound source and the microphone, which reduces the extent to which W-DO is satisfied. In most of the TF points outside the direct path, the corresponding (α, δ) paired parameters cause the peaks in the histogram to deviate from their correct positions at different levels. The level of deviation caused by these TF points cannot be predicted. Moreover, because the activities of sound sources may be continuous or intermittent and mixtures carry noise floors, highlighting peaks becomes difficult; thus, accurately identifying peak positions is challenging. If correct peak positions cannot be found, which prevents the identification of the accurate paired parameters of unknown sound sources, the reconstruction of sound sources is deemed unsatisfactory. Therefore, solving the aforementioned problem and establishing a satisfactory histogram are critical steps in the DUET.

III. PROPOSED METHODS

The histogram defined by the DUET is created by calculating the (α, δ) values corresponding to all the TF points. After the bins corresponding to the (α, δ) values for the α -δ plane is determined, the cumulative number of (α, δ) values for each bin are calculated in the form of a bar height. The number of peaks identified in the histogram represents the number of sound sources, and the bins the peaks belong to are the estimated values of (α, δ). Because of the reverberation caused by intermittent sound sources and the real environment as well as noise disturbance, mixtures do not completely satisfy the W-DO, which renders the peaks difficult to observe and hinders the identification of the real parameters of unknown sources. Therefore, the bar heights in the histogram alone are insufficient for determining $\left(\bar{\alpha}_n, \bar{\delta}_n\right)$, and the short-time data distribution and changes within a window frame must be analyzed. In this study, each time frame in mixtures was designated as the unit time for analyzing signal characteristics in short-time Fourier transform calculations to establish histograms corresponding to each mixture, which are referred to as subhistograms. The discrete time index τ, which describes the TF diagram, is used to describe the subhistograms. On the basis of (7), each subhistogram is expressed as follows:

$$h_{\tau,\Delta\alpha,\Delta\delta}(B, D) = \Sigma_\omega m_{\tau,B,\Delta\alpha}(\omega)\, m_{\tau,D,\Delta\delta}(\omega) \times \left|\hat{x}_{1,\tau}(\omega)\, \hat{x}_{2,\tau}(\omega)\right|^p \omega^q \quad (9)$$

where $m_{\tau,B,\Delta\alpha}$ and $m_{\tau,D,\Delta\delta}$ represent the selection conditions for α and δ, respectively.

$$m_{\tau,B,\Delta\alpha}(\omega) = \begin{cases} 1 & \text{if } |\alpha_\tau(\omega) - B| < \Delta\alpha \\ 0 & \text{otherwise} \end{cases}$$
$$m_{\tau,D,\Delta\delta}(\omega) = \begin{cases} 1 & \text{if } |\delta_\tau(\omega) - D| < \Delta\delta \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where $h_{\tau,\Delta\alpha,\Delta\delta}(B, D)$ represents the subhistogram and B and D represent the bin indices for α and δ, respectively ($B = 1, 2, \ldots, k$; $D = 1, 2, \ldots, l$). In (8) and (10), in case the selected values of $\Delta\alpha$ and $\Delta\delta$ is too small, it may result in histogram not exhibiting all the peaks of sound sources. On the other hand, too large values will make the resolution worse. However, this can be adjusted through empirical attempts. Thus, the subhistogram of each unit time can be observed (Fig. 2). On each subhistogram, irregular patterns of low heights (shown in colors) appear at many positions excluding the peaks. These patterns indicate the biases in the TF points outside the direct path between the sound source and the sensor.

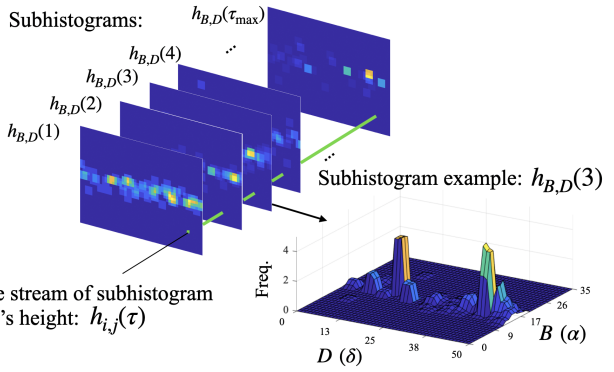


FIGURE 2. Subhistogram $h_{\tau,\Delta\alpha,\Delta\delta}(B, D)$, which is generated using the TF chart of each time frame.

The bar height of the bin of $(B, D) = (i, j)$ can be expressed as $h_{i,j}(\tau)$ in the entire series of subhistograms, where $\tau = 1, 2, \ldots, \tau_{max}$ (i.e., $h_{i,j}(\tau) = h_{\tau,\Delta\alpha,\Delta\delta}(i, j)$).

The analysis results indicate that a visual plane can be established using the infinity norm for the easy observation of intermittent sound sources and the accurate calculation of the number of unknown sources. Moreover, a visual plane based on variances can be constructed to highlight the peaks and reinforce the robustness against noise floors. The remainder of this section describes the new methods proposed in this paper.

A. MAXIMUM DISTRIBUTION

The bar heights of sound sources with high activity time percentages tend to be significantly higher than those of instantaneous sound sources in a histogram. Thus, instantaneous sound sources are difficult to detect and easy to overlook, which leads to source separation failure. Regardless of the activity time percentage, when a sound source is active, a significant bar height is generated in its subhistogram. Accordingly, the highest instantaneous bar height in a subhistogram should be preserved to highlight the presence of an instantaneous sound source. In this study, the infinity norm was used to capture the maximal $h_{i,j}(\tau)$ value, which was used to obtain α and δ height distribution on the plane. This distribution is referred to as the distribution of maximum bar height, which is also known as the maximum distribution, or Max-

Dist. Let a random variable $X_{i,j}(\tau)$ denote the $h_{i,j}(\tau)$ value that occurs in the τth subhistogram ($\tau = 1, 2, \ldots, \tau_{max}$). In this case, Max-Dist is defined as follows:

$$P_{max}(B, D) = \left[||X_{i,j(\tau)}||_{\infty}\right]_{k\times l} \tag{11}$$

where $P_{max}(B, D)$ represents Max-Dist; B and D represent the bin indices for α and δ, respectively ($B = 1, 2, \ldots, k$; $D = 1, 2, \ldots, l$); and $||X_{i,j}(\tau)||_{\infty}$ represents the maximal value of $X_{i,j}(\tau)$. Compared with histograms, Max-Dist considerably reduces the peak differences caused by the time percentage differences between sound sources, thereby enabling easy detection of instantaneous sound sources. The experimental results of this study confirmed that Max-Dist is a suitable index for observing peaks and estimating the correct number of unknown sound sources and accurate peak positions.

B. VARIANCE DISTRIBUTION

Ideally, when a microphone is used to receive a full-bandwidth sound source that has been active for a long time through only a direct path, only a slender peak appears in the subhistogram of every instant. Moreover, the peak's position does not change over time. With disturbances by reflections or reverberations in the real environment, unpredictable deviations occur in the peak positions. However, the instantaneous biases are highly likely to neighbor the accurate position $(\bar{\alpha}_n, \bar{\delta}_n)$ closely and usually occur around it. This phenomenon indirectly leads to frequent and massive changes in the bar height of the accurate position over time. Thus, the position of each $(\bar{\alpha}_n, \bar{\delta}_n)$ point on a subhistogram is the least stable position in terms of bar height changes. Therefore, changes in bar heights are a critical indicator of accurate peak positions. Larger changes in $h_{i,j}(\tau)$ lead to a higher likelihood for (i, j) to be the accurate position of an unknown sound source. Conversely, the data distribution of $h_{i,j}(\tau)$ farther from the accurate source position is more stable and closer to zero (see Fig. 3 for the numerical analysis).

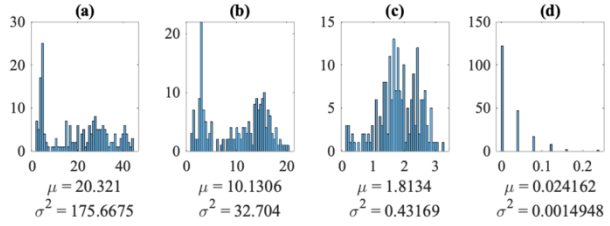


FIGURE 3. Distribution chart, means, and variances of $h_{i,j}(\tau)$, where the x-axis represents $h_{i,j}$ and the y-axis represents the occurrence frequency of $h_{i,j}$: (a) when (i, j) is at the top of the peak, (b) when (i, j) is at the middle of the peak, (c) when (i, j) is at the bottom of the peak, and (d) when (i, j) is far away from the peak. The closer (i, j) is to the top of the peak, the higher the variance becomes. At positions far away from the peak, nearly no changes are detected in the bar heights and the variance is close to 0.

Let $X_{i,j}$ be a random variable with various values of $h_{i,j}(\tau)$, $\tau = 1, 2, \ldots, \tau_{max}$. Let $\mu_{i,j}$ denote the mean of $X_{i,j}$. Because a histogram is the accumulation of $h_{i,j}(\tau)$ (i.e., $H_{\Delta\alpha,\Delta\delta}(B, D) = \Sigma_{\tau}\left[h_{i,j}(\tau)\right]_{k\times l}$), $\mu_{i,j}$ represents the first moment of $X_{i,j}$ (i.e., the mean). The following equation indicates that the accumulation of $h_{i,j}(\tau)$ and the mean of $X_{i,j}$ are proportional:

$$\Sigma_{\tau}\left[h_{i,j}(\tau)\right]_{k\times l} \propto \left[\mu_{i,j}\right]_{k\times l} \tag{12}$$

Thus, histograms represent the first moments of bar heights. If $(\bar{\alpha}_n, \bar{\delta}_n)$ is situated in the (i, j) bin, the queue data on the bin have a higher mean than those in the surrounding areas, which is the basis of the original histogram defined using the DUET. Let the second central moment (i.e., variance) of $h_{i,j}(\tau)$ be the new metric index for the bar height distribution on the $\alpha\delta$ plane. This index is referred to as the distribution of bar height variance and is also known as variance distribution, or Var-Dist. Var-Dist is defined as follows:

$$P_{var}(B, D) = \left[\sigma^2_{i,j}\right]_{k\times l} \tag{13}$$

where $P_{var}(B, D)$ represents Var-Dist and $\sigma^2_{i,j} = \text{var}\left[X_{i,j}(\tau)\right]_{k\times l} = E\left[\left(X_{i,j}(\tau) - \mu_{i,j}\right)^2\right]_{k\times l}$. The closer (i, j) is to the accurate position of the unknown sound source, the higher is the variance in $h_{i,j}(\tau)$, as confirmed by the variance comparison depicted in Fig. 3. The peak positions in the Var-Dist diagram indicate the correct paired attenuation–delay parameters of unknown sound sources.

Notably, the higher the number of TF points that do not satisfy the W-DO in a sample, the fewer is the number of TF points with accurate values of $(\bar{\alpha}_n, \bar{\delta}_n)$. Thus, the higher the number of TF points that do not satisfy the W-DO in a sample, the shorter is the bar height of the accurate peak position. Reduced bar heights are also observed in the surrounding area of the accurate peak position. In the extreme case, the peak may deviate from its accurate position. The changes in the bar heights of the accurate peak position and its surrounding area caused by the weakened W-DO in the Var-Dist diagram are not as profound as those in the original histogram. The reasons for this phenomenon are as follows: (i) the changes in $h_{i,j}(\tau)$ at the accurate position are less stable than those in its surrounding area and (ii) the variance $\sigma^2_{i,j}$ of the accurate position is higher than that at any point in the surrounding area. The differences in variances between the aforementioned positions enable easy identification of the accurate peak position of an unknown sound source (Fig. 4).

IV. EXPERIMENTAL RESULTS

Two examples are presented to demonstrate the advantages of the methods employed in this study over the conventional histogram. The first example involves using Max-Dist to test the changes in the difference in the bar heights of the peaks between the two sound sources with different time percentages. The second example demonstrates the improvements of Var-Dist over the conventional histogram in terms of

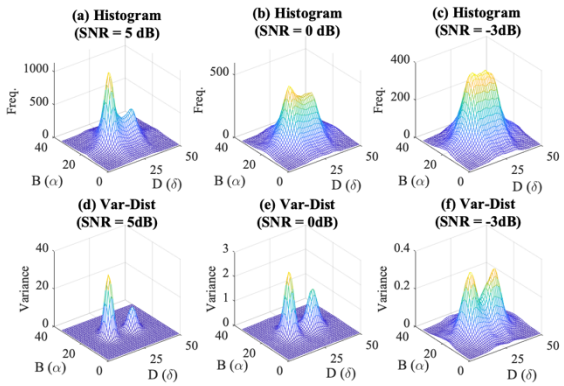


FIGURE 4. Comparison between histograms and Var-Dist in terms of the effectiveness of analysis on samples with white noise: (a) histogram analysis with a signal-to-noise ratio (SNR) of 20 dB, (b) histogram analysis with an SNR of 5 dB, (c) histogram analysis with an SNR of −3 dB, (d) Var-Dist analysis with an SNR of 20 dB, (e) Var-Dist analysis with an SNR of 5 dB, and (f) Var-Dist analysis with an SNR of −3 dB. The shapes of the peaks in the Var-Dist analysis are considerably improved over those in the histograms.

the deviation of peaks from their accurate positions and the shapes of the peaks.

A. EXAMPLE 1: MAXIMUM DISTRIBUTION

Example 1 indicates that the peak height difference between two sound sources as indicated by Max-Dist is not significantly affected by the difference in their activity time percentages. An experiment was conducted in an indoor environment, where audio samples were recorded at a sampling frequency of 16,000 Hz according to the restriction condition listed in (6). Ideally, the two microphones must be set no farther apart than 2.15 cm from each other. The two microphones were set 1.5 cm apart from each other in this study to fulfill the aforementioned condition. Speakers were set at a distance 1m and at 45° and 135° from each microphone to broadcast music and speech simultaneously for forming mixtures.

After the recording was complete, the mixtures were cropped into two 40,000-point segments in length to compare the analysis results for different time percentage differences between the two sound sources. The frequencies of the two segments were adjusted and merged to create mixtures with different time percentages from the two sound sources. The segment x_k^{s} contained only music as the active sound source, and this segment was considered to represent a sound source with a high percentage of activity time. The segment x_k^c comprised music and speech as active sources, and this segment was considered to represent instantaneous sound source activities. The aforementioned two segments were arranged in various proportions to create new mixtures by using the following "repmat" function from MATLAB:

for $k \leftarrow 1$ **to** 2 **do**
$x_k^{\mathrm{mix}} \leftarrow [\mathbf{repmat}(x_k^c, L, 1); \mathbf{repmat}(x_k^{\mathrm{s}}, 50 - L, 1)];$
end for

where x_k^{mix} represents the kth new mixture ($k = 1, 2, \ldots$) and $L = 0, 1, 2, \ldots, 50$ represents the number of which the instantaneous source x_k^c is repeated. Each mixture was created by repeating $x_k^c L$ times and then repeating $x_k^{\mathrm{s}} 50 - L$ times in succession. The total length of each completed mixture is 2,000,000 points. Thus, the mixtures required in this study were created simply by adjusting L, and the time percentage of the instantaneous sound source constituted $2L\%$ of the total length of the mixture. Fig. 5 illustrates the recorded samples in this study, and source separation was conducted by adopting the DUET after obtaining Max-Dist by using (11). Fig. 6 presents a comparison of the peak changes in the conventional histogram and Max-Dist.

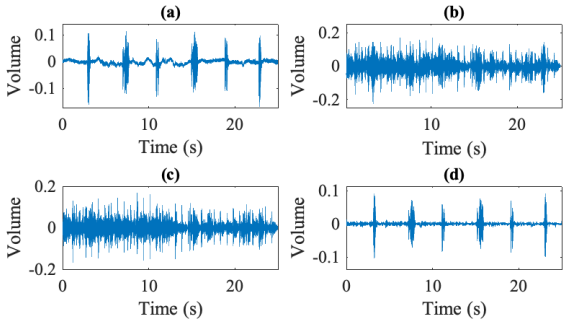


FIGURE 5. Recorded samples in this study and source separation through the DUET: (a) clean speech source; (b) one of the mixtures; (c) separated music source, which exhibited a high time percentage; and (d) separated speech source, which exhibited a low time percentage.

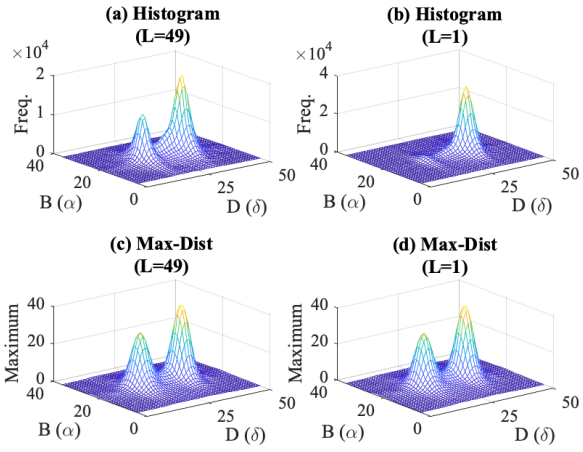


FIGURE 6. Comparison between the histogram and Max-Dist on peak values: (a) histogram with x_k^c constituting 98% of the total mixture length, (b) histogram with x_k^c constituting 2% of the total mixture length, (c) Max-Dist with x_k^c constituting 98% of the total mixture length, and (d) Max-Dist with x_k^c constituting 2% of the total mixture length. As shown in (c) and (d), no significant changes occurred in Max-Dist. Moreover, the peaks in Max-Dist correspond to the speech source, which was shorter but more active than the music source.

The peak height difference ratio d_{peak} between the two sound sources is defined as the bar height difference as follows:

$$d_{peak} = (h_{\mathrm{peak1}} - h_{\mathrm{peak2}})/h_{\mathrm{peak1}} \tag{14}$$

where h_{peak1} and h_{peak2} denote the high peak and low peak, respectively. The ratio can be regarded as an index of the quality of the performed experiment. The lower the ratio, the higher was the satisfaction level for the experiment. Fig. 7 presents a comparison between the conventional histogram and Max-Dist method.

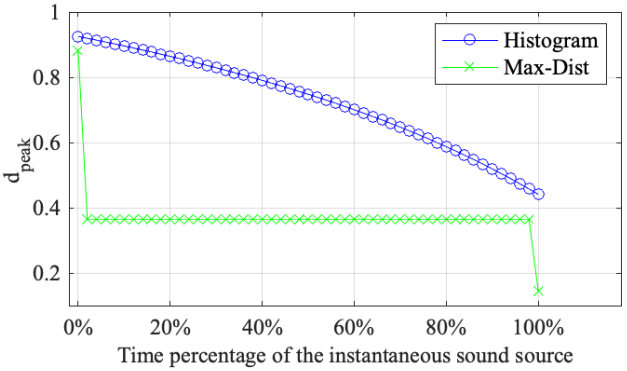


FIGURE 7. Comparison between the histogram and Max-Dist on the peak height difference ratios (d_{peak}) with various time percentages for x_k^c. A higher time percentage for x_k^c led to a lower d_{peak} value in the histogram. In Max-Dist, d_{peak} decreased to 0.3650 when the time percentage of x_k^c started to increase from 0% and remained consistent until the time percentage reached 100%.

As displayed in Fig. 7, the d_{peak} in the histogram decreased gradually as the time percentage increased for x_k^c. When the time percentage of x_k^c was 2%–98% in Max-Dist, the peak values of both sound sources were maintained at the maximal height; thus, d_{peak} remained constant. Although d_{peak} changed when the time percentage reached 100% because of a lack of information from one of the two segments in the mixtures, the d_{peak} value in Max-Dist was considerably lower than that in the histogram. Therefore, the time percentages of the sound sources in the histogram significantly influenced the peak heights. A higher time percentage difference between the sources led to a higher difference ratio in their peak heights. The peak height difference ratio in Max-Dist was up to 50% lower than that in the histogram.

Thus, the presence of a sound source can be highlighted by Max-Dist regardless of its time percentage. Compared with the histogram, Max-Dist enables more effective observation of the peaks of instantaneous sound sources, which leads to the reliable calculation of the number of unknown sound sources in mixtures.

B. EXAMPLE 2: VARIANCE DISTRIBUTION

Example 2 demonstrates the advantage of Var-Dist over the conventional histogram in accurately estimating peak positions under a high noise disturbance. The test environment in Example 2 was the same as that used in Example 1; however, the music source in Example 1 was replaced with a speech source with combined male and female voices in Example 2. The original speech source was the same in both examples.

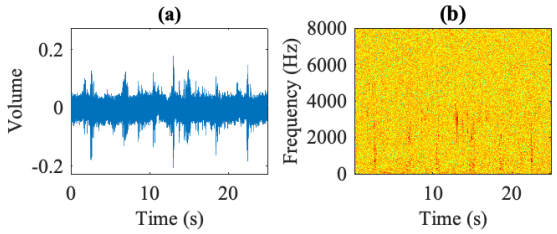


FIGURE 8. (a) A waveform of mixture with an SNR of 0 dB and (b) high noise disturbance in the TF chart.

After the recording was complete, the two obtained mixtures were mixed with white noise to weaken the W-DO of the sound sources substantially (Fig. 8). The noise disturbance was adjusted using the signal-to-noise ratio (SNR). A smaller SNR indicated higher noise disturbance. The SNR is defined as follows:

$$SNR := 10\,log_{10}\frac{||x_k + e_k||^2}{||e_k||} \tag{15}$$

where x_k and e_k represent the kth mixture and the random white noise it carries ($k = 1, 2, \ldots$), respectively. The peak value was defined using the following equation to estimate the bias in the peak position for determining the quality of the conducted experiment:

$$d_{bias} = \frac{1}{N}\sum_n^N \sqrt{(\alpha_n - \bar{\alpha}_n)^2 + \left(\delta_n - \bar{\delta}_n\right)^2} \tag{16}$$

where d_{bias} represents the average bias in the estimated peak position when each grid in the (B, D) plane is considered a unit. A smaller d_{bias} value indicates more accurate estimation results. The total number of bins for both axes was set as 125. The parameters $\bar{\alpha}_n$ and $\bar{\delta}_n$ represent the correct values for the n^{th} sound source, determined by the attenuation and delay obtained from the histogram prior to adding noise, whereas α_n and δ_n indicate the estimated values for this sound source. The conventional histogram and Var-Dist, which was established using (13), were compared in terms of the biases between the estimated and correct peak positions under SNRs between 20 and -20 dB. At each SNR, random noise was tested 100 times (see the top diagram in Fig. 9). Ideally, a peak should be tall and slender. To compare the peak shapes generated using the conventional histogram and Var-Dist, the full width at half maximum (FWHM) standard was employed as the index of the ideal peak shape. A lower FWHM value indicated a more suitable peak shape (see the bottom diagram in Fig. 9).

As displayed in Fig. 9, among the three adopted methods (i.e., histogram, Var-Dist, and Max-Dist), Var-Dist exhibited the lowest peak position biases and the slenderest peak shapes at most SNRs. Although Max-Dist and Var-Dist did not differ substantially in position biases, the peaks obtained with Max-Dist were considerably wider than those obtained with Var-Dist. Therefore, the peaks obtained with Var-Dist were the most stable peaks at most SNRs. The difference in peak position biases between Var-Dist and the histogram began

TABLE 1. Comparison of different methods for determining attenuation and delay.

	Related terms in mathematics	Unit models	Intention
Histogram	The first moment	$\left[\mu_{i,j}\right]_{k\times l}$	The histogram is a classic method for determining the peaks of sound sources. Peak heights are directly influenced by the activity time percentages of sound sources. When mixtures are disturbed by noise, peaks become unnoticeable and severely biased in certain positions.
Max-Dist	Infinity norm	$\left[\|h_{i,j}(\tau)\|_\infty\right]_{k\times l}$	Max-Dist allows the number of sound Sources to be found easily. It highlights the peak heights of sound sources with short activity time percentages.
Var-Dist	The second central moment	$\left[\sigma^2_{i,j}\right]_{k\times l}$	Var-Dist is used under low SNRs. It presents peak shapes that are easy to observe and less biased than those in the histogram in most cases.

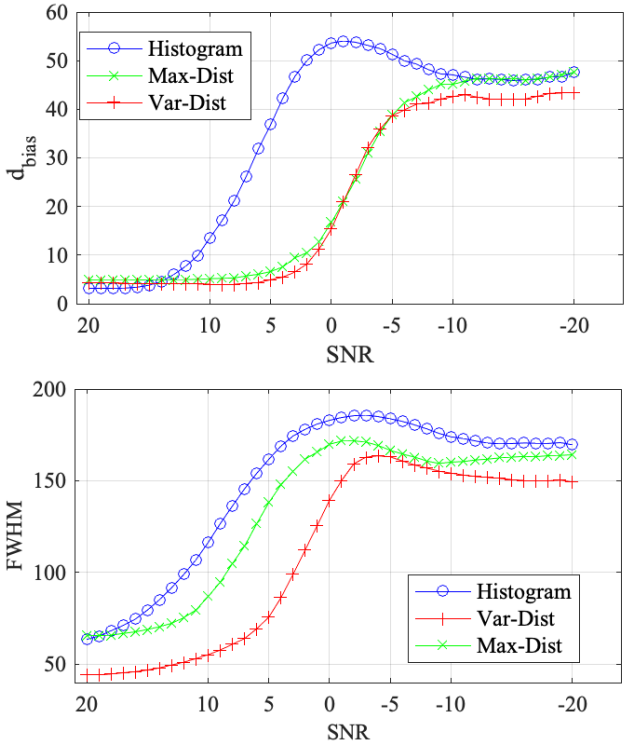


FIGURE 9. Biases in the estimated peak positions (top) and the FWHM indices (bottom) under SNRs between 20 and −20 dB.

to decrease gradually only after the SNR reached −3 dB or lower. Compared with histogram, Var-Dist reduces the bias by 86.6% when SNR is 5dB, reduces by 71.32% when it is 0dB, and reduces by 39.5% when it is -3dB. According to [11], the weight parameters ($p = 2$, $q = 2$) in the histogram of the DUET can be adjusted to improve the peak estimation accuracy at a low SNR. This finding is consistent with the results of the experiment conducted for Example 2. Fig. 10 depicts a comparison between the three analysis methods under the same condition.

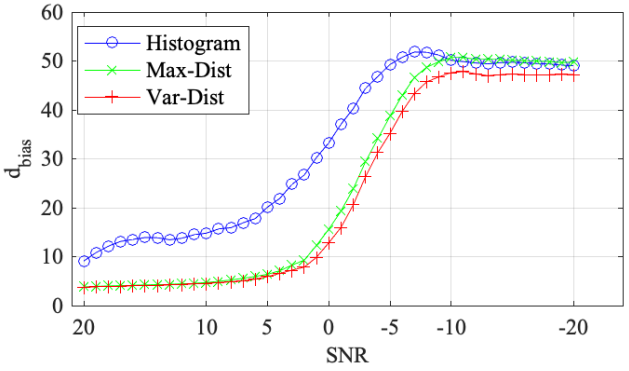


FIGURE 10. Peak position biases at the weight parameters of p = 2 and q = 2 under white noise disturbance for SNRs between 20 and −20 dB.

As displayed in Fig. 10, because of the weight parameters p and q, the biases in the histogram decreased substantially when SNR ranged between −5 and 5 dB. However, under the same condition Var-Dist exhibited the lowest biases. When the SNR reached −10 dB or below, the noise had completely obstructed the true data and the bias differences between the three methods ceased to be significant. The results of the aforementioned experiment confirmed that the biases in Var-Dist remained extremely low even when the W-DO in the mixtures was weak. This result indicates the strong robustness of Var-Dist against noise disturbance and the ease of observing peak shapes by using this index. Therefore, Var-Dist is more suitable than the conventional histogram for analyzing peaks in mixtures with a low SNR. According to the results, Table 1 presents a comparison of these two indices and the conventional histogram.

C. ACTUAL CASES

Here we used real recordings that have not been edited or processed as experimental mixtures and considered two cases to reveal that Max-Dist and Var-Dist proposed in this work have indeed enhanced the performance of the conventional

DUET histogram. In terms of experimental environment settings, the indoor space, the distance between microphones, and the angle and distance of the sound sources are the same as Example 1 and Example 2.

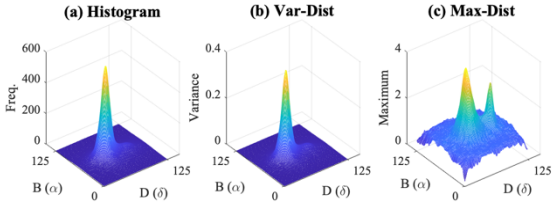


FIGURE 11. Patterns of peaks in Case 1 using: (a) Histogram, (b) Var-Dist, and (c) Max-Dist. It can be seen that (a) and (b) cannot capture the peaks of the instantaneous sound source.

1) CASE 1: MAX-DIST

The two sound sources are male and female speeches. The total time of the mixtures is 60 seconds. The female speech time is quite short, only 3.6 seconds in total. Experimental results proved that neither the DUET histogram nor the Var-Dist method can determine the attenuation and delay of mixtures, only Max-Dist can do it. As shown in Fig. 11, only one peak is displayed on the histogram and Var-Dist, while Max-Dist can highlight the peak positions of the two sound sources. We then separated two speech sources based on the attenuation delay obtained by Max-Dist, which is shown in Fig. 12. As shown in Fig. 11, only one peak is displayed by the histogram and Var-Dist, while Max-Dist can highlight the peak positions of the two sound sources. We then separated these two speech sources based on the attenuation and delay obtained by Max-Dist, which are shown in Fig. 12.

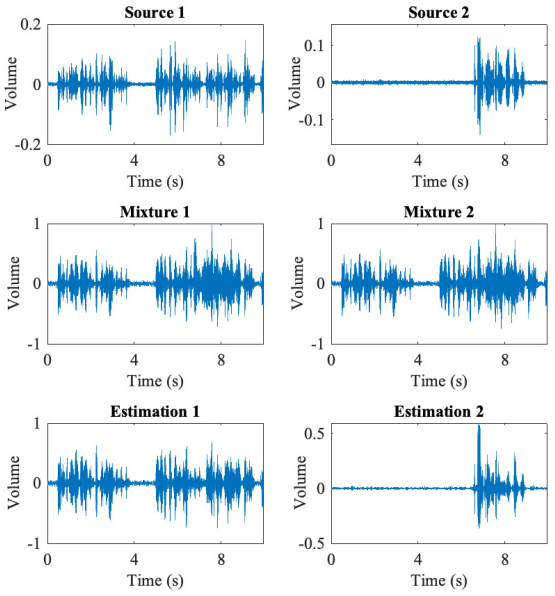


FIGURE 12. Comparison of waveforms before and after sound source separation. In order to check the correctness of the sound sources separated by the Max-Dist method, only the partial time length is shown here.

In this case, the sound source with short time percentage results in too small peak on the conventional histogram, which is unable to be determined. However, the Max-Dist method can completely overcome this problem.

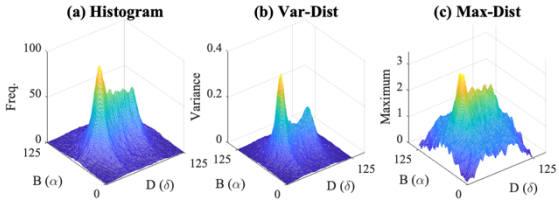


FIGURE 13. Patterns of peaks in Case 2 using: (a) Histogram, (b) Var-Dist, and (c) Max-Dist. It can be seen that the second peak corresponding to (a) and (c) is very ambiguous, which will lead to incorrect subsequent separation.

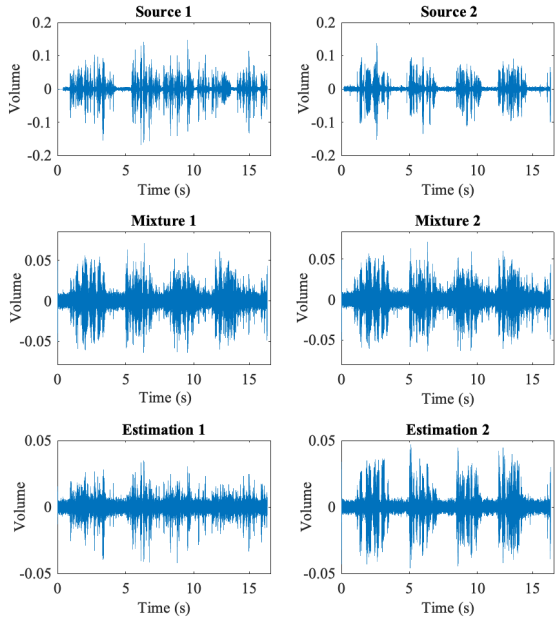


FIGURE 14. The waveform diagrams in Case 2 contain the sound sources (not yet polluted by noise), mixtures, and the estimated sound sources separated by Var-Dist. Among them, source 2 is the sound source corresponding to the second peak discussed in Fig. 13. In order to avoid unpredictable factors in the three methods, the separation results without noise reduction were compared.

2) CASE 2: VAR-DIST

In this Case, we set up two loudspeakers on the ceiling to play rain recordings as environmental sound effects, corresponding to the white noise simulated in Example 2. The two sources are male and female speeches with similar activity time percentages. The experimental results are shown in Fig. 13, where only one peak of histogram and Max-Dist can be confirmed, and the peak corresponding to the other sound source is difficult to identify where it is located. Var-Dist can present two obvious peaks. Based on the attenuation and delay obtained by the Var-Dist method, we subsequently separated two noise-contained sound sources, and showed

the sound sources and mixtures used in this experiment in Fig. 14.

This real case shows that a certain degree of noise interference will make the histogram unable to determine the correct peak position of the sound source. Var-Dist can overcome this problem. For mixtures with excessive noise, without applying any noise reduction, Var-Dist can obtain the correct parameters for subsequent separation of DUET.

In the end, we fed the mixtures in Case 1 and 2 as the data to other well-known methods for comparison. The performance indices for comparison are SDR and SIR. The tool we used is the well-known BSS_EVAL_TOOL [22]. The methods included in the comparison are listed below: (i) Dictionary Learning Blind Sources Separation (DL) [23], (ii) Fast Fixed point ICA (FastICA) [5], (iii) Equivariant Robust Independent Component Analysis algorithm (ERICA) [24], (iv) Convolutive blind source separation (ConvBSS), (v) InfoMAX, (vi) P-ICA[7], (vii) UCBSS: Underdetermined Convolutive Blind Source Separation (UCBSS) [9], (viii) DUET, and (ix) DUET using the proposed Maximum Distribution (Max-D) or the proposed Variance Distribution (Var-D).

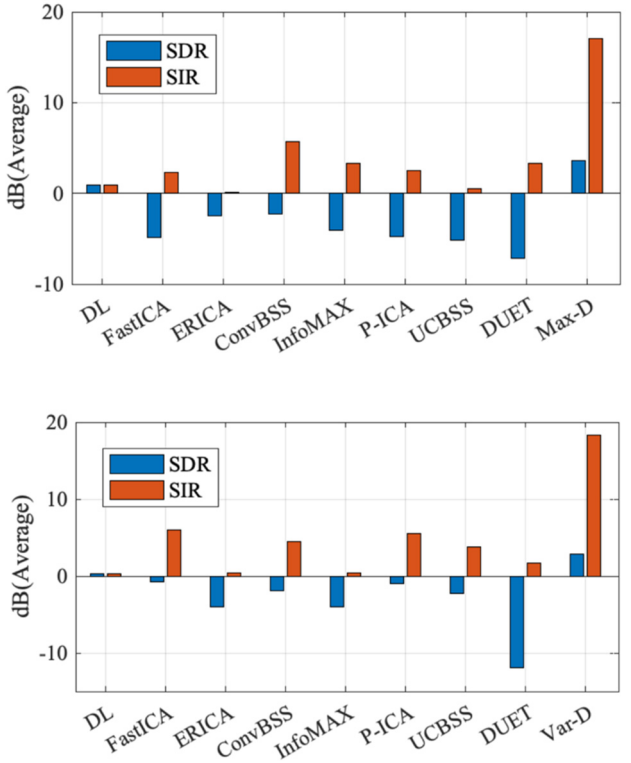


FIGURE 15. Comparison results of various methods using SDR and SIR. (Above) Case 1: The mixtures are the combination of the speeches with long and short activity time percentages. (Below) Case 2: The mixtures are disturbed by excessive noise. Max-Dist and Var-Dist can correctly find the attenuation and delay of the sound sources and use DUET to decompose them.

The comparison results in Fig. 15 show that the performance of our proposed methods is excellent. Other methods cannot effectively deal with Case 1 and 2, because such special circumstances will strongly affect the assumptions of those methods on the distribution of sound sources.

V. CONCLUSION

In BSS through the DUET, efficiently determining peak positions in the histogram is critical for reconstructing sound sources. An unsatisfactory histogram prevents the accurate identification of parameters corresponding to unknown sound sources, which decreases the likelihood of successful source separation. Realistically, some sound sources are difficult to detect in a histogram because of their short time percentages, and excessive noise floors in mixtures also prevent the confirmation of peaks. Therefore, the conventional histogram was converted into Max-Dist and Var-Dist in the current study to solve the aforementioned problems. The experimental results confirmed that Max-Dist and Var-Dist were superior to the conventional histogram in terms of the peak shapes and peak position estimation accuracy. At a low SNR, Var-Dist exhibited the strongest robustness among the three methods. In most cases, Max-Dist and Var-Dist can replace the conventional histogram in the direct estimation of the attenuation and delay of unknown sound sources.

REFERENCES

[1] S. Choi, A. Cichocki, H. M. Park, and S.-Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Inf. Process. Lett. Rev.*, vol. 6, no. 1, pp. 1–57, 2005.

[2] S. A. Cruces-Alvarez, A. Cichocki, and S. Amari, "From blind signal extraction to blind instantaneous signal separation: Criteria, algorithms, and stability," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 859–873, Jul. 2004.

[3] P. Common and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. San Diego, CA, USA: Academic, 2010.

[4] Z. Koldovský and P. Tichavský, "Time-domain blind audio source separation using advanced ICA methods," in *Proc. Interspeech*, Aug. 2007, pp. 846–849.

[5] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Syst.*, vol. 10, no. 1, pp. 1–8, Feb. 2000.

[6] Q. Jin, G. Wang, and Y. Liu, "Blind signal separation by entropy maximization (INFOMAX)," in *Proc. Int. Conf. Comput. Intell. Softw. Eng.*, Sep. 2010, pp. 1–5.

[7] K. Zhang and L. W. Chan, "ICA by PCA approach: Relating higher-order statistics to second-order moments," in *Proc. Int. Conf. Independ. Compon. Anal. Signal Separat.* Berlin, Germany: Springer, 2006, pp. 311–318, 2006.

[8] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2007, pp. 139–142.

[9] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined convolutive blind source separation via time–frequency masking," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 101–116, Jan. 2010.

[10] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[11] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation. Signals and Communication Technology*. Berlin, Germany: Springer, 2007, pp. 217–241.

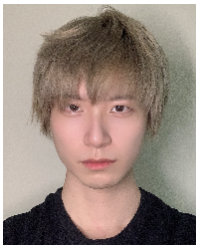
[12] S. M. Abdulla and J. Jayakumari, "DUET using automatic peak detection and histogram thresholding," in *Proc. IEEE Int. Conf. Circuits Syst. (ICCS)*, Dec. 2017, pp. 291–296.

[13] A. A. S. Gunawan, A. Stevelino, H. Ngarianto, W. Budiharto, and R. Wongso, "Implementation of blind speech separation for intelligent humanoid robot using DUET method," *Proc. Comput. Sci.*, vol. 116, pp. 87–98, Jan. 2017.
[14] A. Kumar, A. Mukherjee, and M. Mandava, "Estimation of speed and tracking of vehicles using radar duet," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2019, pp. 1–5.
[15] H. Pu, C. Cai, M. Hu, T. Deng, R. Zheng, and J. Luo, "Towards robust multiple blind source localization using source separation and beamforming," *Sensors*, vol. 21, no. 2, p. 532, Jan. 2021.
[16] K. J. Faller, J. Riddley, and E. Grubbs, "Automatic blind source separation of speech sources in an auditory scene," in *Proc. 51st Asilomar Conf. Signals, Syst., Comput.*, Oct. 2017, pp. 248–250.
[17] V. Mordoh and Y. Zigel, "Audio source separation to reduce sleeping partner sounds: A simulation study," *Physiolog. Meas.*, vol. 42, no. 6, Jun. 2021, Art. no. 064004.
[18] R. Balan, J. Rosca, S. Rickard, and J. O'Ruanaidh, "The influence of windowing on time delay estimates," in *Proc. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2000, pp. 15–17.
[19] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jun. 2000, pp. 2985–2988.
[20] Y. Wang, Ö. Yılmaz, and Z. Zhou, "Phase aliasing correction for robust blind source separation using DUET," *Appl. Comput. Harmon. Anal.*, vol. 35, no. 2, pp. 341–349, Sep. 2013.
[21] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[22] A. Ciaramella, D. Nardone, and A. Staiano, "Blind source separation using dictionary learning in wireless sensor network scenario," in *Neural Approaches to Dynamics of Signal Exchanges*. Singapore: Springer, 2020, pp. 119–131.
[23] P. P. Singh and R. D. Garg, "Classification of high resolution satellite images using equivariant robust independent component analysis," in *Advanced Computing, Networking and Informatics*, vol. 1. Cham, Switzerland: Springer, 2014, pp. 283–290.

KUANG-YOW LIAN (Member, IEEE) received the B.S. degree in engineering science from the National Cheng Kung University, Tainan, Taiwan, in 1984, and the Ph.D. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1993.

From 1994 to 2007, he was an Associate Professor, a Professor, and the Chair of the Department of Electrical Engineering, Chung Yuan Christian University, Zhongli, Taiwan. He is currently a Professor with the Department of Electrical Engineering, National Taipei University of Technology, where he also worked as the Chair, from 2009 to 2012. His research interests include smart sensor technology, smart living devices, machine learning, robotics, and control system applications.

Prof. Lian has received the following honors and awards: the Chinese Automatic Control Society (CACS) Fellow in 2015, the CACS Outstanding Automatic Control Engineering Award in 2012, and the 2014 and 2017 Macronix Gold Silicon Best Advisor Awards.



JIA-HSIN LIN received the M.Sc. degree in electrical engineering from Chung Yuan Christian University, Taoyuan City, Taiwan, in 2012. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan. His main research interests include blind sources separation, non-negative matrix factorization, and musical instrument recognition.