



Dual transform based joint learning single channel speech separation using generative joint dictionary learning

Md Imran Hossain¹ · Tarek Hasan Al Mahmud² · Md Shohidul Islam³ ·
Md Bipul Hossen¹ · Rashid Khan¹ · Zhongfu Ye¹

Received: 29 September 2020 / Revised: 21 January 2022 / Accepted: 9 March 2022 /
Published online: 2 April 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Single channel speech separation (SS) is highly significant in many real-world speech processing applications such as hearing aids, automatic speech recognition, control humanoid robots, and cocktail-party issues. The performance of the SS is crucial for these applications, but better accuracy has yet to be developed. Some researchers have tried to separate speech using only the magnitude part, and some are tried to solve complex domains. We propose a dual transform SS method that serially uses the dual-tree complex wavelet transform (DTCWT) and short-term Fourier transform (STFT), and jointly learns the magnitude, real and imaginary parts of the signal applying a generative joint dictionary learning (GJDL). At first, the time-domain speech signal is decomposed by DTCWT, which produces a set of subband signals. Then STFT is connected to each subband signal, which converts each subband signal to the time-frequency domain and builds a complex spectrogram that prepares three parts like real, imaginary and magnitude for each subband signal. Next, we utilize the GJDL approach for making the joint dictionaries, and then the batch least angle regression with a coherence criterion (LARC) algorithm is used for sparse coding. Afterward, computes the initially estimated signals in two different ways, one by considering only the magnitude part and another by considering real and imaginary components. Finally, we apply the Gini index (GI) to the initially estimated signals to achieve better accuracy. The proposed algorithm demonstrates the best performance in all considered evaluation metrics compared to the mentioned algorithms.

Keywords Speech separation (SS) · Dual-tree complex wavelet transform (DTCWT) · Generative joint dictionary learning (GJDL) · Short-time Fourier transform (STFT) · Gini index (GI)

✉ Zhongfu Ye
yezf@ustc.edu.cn

1 Introduction

Speech separation is a process where several signals have been combined together, and the goal is to retrieve the original signals from the mixed signal. Speech separation has pulled in a striking measure of consideration because of its expected use in several real-world applications, for example, hearing aids, automatic speech recognition, communication, medical, multimedia, assisted living systems, control humanoid robots, cocktail-party issue, and so forth [9, 29, 32, 47]. In these applications, well-separated signals are obligatory for the system to work appropriately. According to the number of channels, the speech separation problem is categorized into multichannel, binaural channel, and single channel types. A single-channel speech separation (SCSS) process [4, 17, 26], which still remains a significant research challenge because only one recording is available, and the spatial information that can be extracted is restricted [34].

With the current growing interest in speech separation, many SCSS models have been proposed in considering numerous parameters, for example, phase, magnitude, frequency, energy, and the spectrogram of the speech signal. The factorial hidden Markov models (HMMs) have been suggested by S.T. Roweis that are incredibly successful in demonstrating a single speaker [31]. Jang and Lee [15] use a maximum likelihood approach to separate the mixed source signal that is perceived in a single channel (SC). Researchers increasingly use nonnegative matrix factorization (NMF) to separate SC source signals. It's a collection of methods in a multivariate analysis where a matrix is decomposed into two other nonnegative matrices according to its components and weights. NMF was first presented by Paatero and Tapper [22] and emerged for the use of source separation by Lee and Seung [28]. Sparse nonnegative matrix factorization (SNMF) is applied to factorize the speech signals [41, 45]. They used SNMF to learn the sparse representation of the data that solves the problem of separating multiple speech sources from a single microphone recording. Sparsity is prescribed only for signal detection in the coefficient matrix [44].

Recently, wavelet-based separation methods [11, 12, 27, 42] have been emerged to the researchers. In [42], discrete wavelet transform (DWT) and SNMF based speech separation methods is implemented. The authors used the wavelet decomposition to speed up the separation process by rejecting the high-frequency components of the source signals. Here DWT is used that splits a signal into its low-frequency parts known as the approximations coefficients and high-frequency parts known as details coefficients. Though it reduces the separation time, but the intelligibility of individual speakers is severely affected due to the total rejection of high-frequency components. In [11], the stationary wavelet transform (SWT) and NMF based speech enhancement (SE) method is presented. The SWT discards the downsampling approach used in both the DWT and discrete wavelet packet transform (DWPT) at every level to acquire the shift-invariance property. This method leads to redundant problems and cannot use the sparseness among different speech signals. The DTCWT is utilized for speech enhancement by considering the NMF but ignoring the sparsity [12]. Therefore, the assessed speech has become inaudible because a few errors occur during deteriorations of the signal utilizing NMF, i.e., some noises or artifacts have been incited during disintegrations through NMF.

The dictionary learning (DL) algorithm [2, 7, 23, 25, 35, 36, 48, 49] is another useful technique for model-based SCSS. They assume that speech signals that have sparse representations from different speakers have some individual components. Usually speaking, a joint dictionary is a redundant dictionary. One source replies to categorize subdictionaries with

additional sources that can't be avoided, though sparse constraints are applied to train the dictionaries. Some approaches for learning the discriminative dictionary have been improved in the past, such as the Metaface learning method [49]. A series of approaches are presented in [2, 7, 23], where the joint discriminative dictionary is made by varying objective functions or adding penalty items to learn the dictionary. However, the solution of the optimization problem comes to be difficult for the complexity of the objective function, and then the time complexity becomes larger. The adaptive discriminative dictionary learning (ADDL) procedure [2] undertakes that different speakers' speech signals have distinctive constituents. The dictionary column is known as an atom that is rational to the signal if the absolute value of the inner product of the atom and the signal is large. Subsequently, using a discriminative dictionary to code the several organized speech signals sparsely, the coding coefficients of the different essential sources are disseminated separately over all dictionary elements. Sequential discriminative dictionary learning (SDDL) is presented in [48], where both distinctive and similar parts of varying speaker signals are considered. The authors [25] present a sparsity model comprising a couple of joint sparse representations (JSR). The mapping relationship between mixture and speech is used in one JSR, while the mapping relation between speech and noise is used in another. The authors of [35] construct a joint dictionary with a common sub-dictionary method (CJD) where a CJD is built using similar atoms between identity sub-dictionaries. The identity sub-dictionaries are trained using source speech signals corresponding to each speaker. In [36], the authors offered a new optimization function to formulate a joint dictionary (OJDL) with multiple identity sub-dictionaries and a common sub-dictionary. The authors of [5] proposed a two-level correlative joint sparse portrayal technique to improve the performance of single-channel speech separation. To suppress noise source confusion, a two level joint inadequate portrayal was built utilizing the relationship between speech, mixture signals, and the discriminative property of joint word dictionary. The authors of [16, 39], proposed a speech improvement technique with substitute advancement of sparse coefficient and dictionary. The Fisher criterion compelled the target capacity of dictionary learning, and afterward, the discriminative dictionary and the sparse comparing coefficient are acquired. Thusly, the irritated obstruction among joint dictionaries can be diminished.

Deep learning has got particular consideration in the SS community in which non-linear mapping among the mixture and speech is considered. Deep learning-based techniques can be divided into two divisions based on the association between the noisy input and desired outputs. They are the deep neural networks (DNN) based mask algorithms [20, 50] and DNN based on regression algorithms [38]. These techniques have been successfully implemented and have shown outstanding performance in improving the desired signal from the mixture signal. Besides, it is not suitable for dealing with limited features, constrained inside the thought of sources, and somewhat greater computational complexity. Therefore, we combined the DTCWT and STFT to take advantage of both transforms and better resolve to process the noisy mixture [8, 13]. Finally, we're applying SNMF after the DTCWT and STFT to get the estimated clean speech. These algorithms work very well; however, only the magnitude spectrum is enhanced and overlooks the phase's enhancement.

Most of the techniques deliberated above consider only the magnitude part, while the phase portion is not enriched. Though the consideration of the magnitude part does have a significant contribution to the estimated speeches, but the improvement of the phase portion cannot be overlooked. The effects of complex estimation have been used in [37, 46], and they found a potential improvement of speech separation.

The contributions of this paper are briefly listed below:

1. For an accurate and in-depth exploration of the signal, we apply dual transforms DTCWT and STFT. The DTCWT decomposes the signal into a set of low and high-frequency component subband signals that makes the signal more stationary. After applying DTCWT, then connected STFT to each subband signal that builds a complex spectrogram for each subband signal. Thus, we better understand the signal for further analyzing and processing.
2. Unlike many other algorithms that investigate either magnitude or phase component, we are handling the magnitude of the signal and the real and imaginary parts. So, these techniques take complete advantage of all the information available in the waveforms of the signals. To accumulate the best version of speech separation, we process the magnitude part, the real part, as well as the imaginary part. As concerns we know, we are the first who jointly investigate the magnitude, the real and imaginary parts of the signal.
3. In this method, we use the DTCWT and STFT consecutively, then apply the GJDL algorithm to both the magnitude part, the absolute value of the real and imaginary part of the signal, and preserve the sign. We apply the LARC algorithm for sparse coding that finds the necessary coefficients. Our suggested approach assesses the initial signals in two separate ways. One is an estimated signal that considers only the magnitude part, and the other is an approximate signal that includes both real and imaginary parts. Finally, the Gini index is used to calculate the complementary impacts of both initially estimated signals. Thus, we get advantages from using the GI with the initially estimated signals.

The rest of the paper is section-wise distributed as follows: A mathematical description of the issue is given in Section 2. Section 3 briefly describes DTCWT, STFT, GDL, and GI. Section 4 is this paper's method, which entires up the functions of each module. Section 5 presents the experimental setting and results using the GRID [30], and TIMIT [53] datasets for speech separation. Finally, section 6 finalizes the paper supplemented with references. The nomenclature is provided in Table 1.

2 Problem formulation

Single-channel speech separation problem can be defined as follows: Assume $\mathbf{z}(t)$ be the mixed signal which includes of two speakers speech signal $\mathbf{x}(t)$ and $\mathbf{y}(t)$. The goal of single channel speech separation is to obtain the estimated signals $\mathbf{x}(t)$ and $\mathbf{y}(t)$ from the mixed signal $\mathbf{z}(t)$. The expression for the mixed signal defines in Eq. (1).

$$\mathbf{z}(t) = \mathbf{x}(t) + \mathbf{y}(t) \quad (1)$$

Now, the DTCWT is applied to Eq. (1) and gets the subbands are presented in Eq. (2) as follows

$$\mathbf{z}_{b,tl} = \mathbf{x}_{b,tl} + \mathbf{y}_{b,tl} \quad (2)$$

where $\mathbf{z}_{b,tl}$, $\mathbf{x}_{b,tl}$ and $\mathbf{y}_{b,tl}$ denote the mixed, first source, and second source subband signals, respectively, and b is one more than the level of DTCWT decomposition, and tl indicates the tree level (describe in subsection 3.1). Now, we apply STFT to each subband signal and time-frequency representation of Eq. (3) can be expressed as

Table 1 Nomenclature

Nomenclature	
x, X (lowercase & uppercase)	Variables
x (lowercase bold)	Vector
X (uppercase bold)	Matrix
<i>X</i> (uppercase italic)	Function
<i>X</i> (uppercase bold italic)	Method
×	Elementwise multiplication
$\sqrt{(\cdot)}$	Elementwise square root operation
SS	Speech separation
STFT	Short-time Fourier transform
ISTFT	Inverse short-time Fourier transform
NMF	Non-negative matrix factorization
SNMF	Sparse non-negative matrix factorization
DWT	Discrete wavelet transform
IDWT	Inverse discrete wavelet transform
GJDL	Generative joint dictionary learning
LARC	Least angle regression with a coherence criterion
SWT	Stationary wavelet transform
DTCWT	Dual-tree complex wavelet transform
IDTCWT	Inverse dual-tree complex wavelet transform
CJD	Joint dictionary method with a common sub-dictionary
OJDL	Optimization joint dictionary learning
GI	Gini index
PM	Proposed method
HASQI	Hearing- Aid’ Speech Quality Index
HASPI	Hearing- Aid’ Speech Perception Index
PESQ	Perceptual evaluation of speech quality
STOI	Short-time objective intelligibility
SDR	Source distortion ratio
SIR	Signal-to-interference ratio

$$\mathbf{Z}_{b,tl}(\tau, f) = \mathbf{X}_{b,tl}(\tau, f) + \mathbf{Y}_{b,tl}(\tau, f) \tag{3}$$

where $\mathbf{Z}_{b,tl}(\tau, f)$, $\mathbf{X}_{b,tl}(\tau, f)$ and $\mathbf{Y}_{b,tl}(\tau, f)$ are the STFT coefficients of $\mathbf{z}_{b,tl}$, $\mathbf{x}_{b,tl}$ and $\mathbf{y}_{b,tl}$, respectively. f and τ represent the frequency bin index and time frame index, respectively.

Eq. (3) can be decomposed the real and imaginary parts as follows, where $\mathbf{ZR}_{b,tl}(\tau, f)$ and $\mathbf{ZI}_{b,tl}(\tau, f)$ represent the real and imaginary parts of $\mathbf{Z}_{b,tl}(\tau, f)$ and the same for others.

$$\mathbf{ZR}_{b,tl}(\tau, f) + i\mathbf{ZI}_{b,tl}(\tau, f) = \mathbf{XR}_{b,tl}(\tau, f) + i\mathbf{XI}_{b,tl}(\tau, f) + \mathbf{YR}_{b,tl}(\tau, f) + i\mathbf{YI}_{b,tl}(\tau, f) \tag{4}$$

The magnitude $|\mathbf{Z}_{b,tl}(\tau, f)|$, real $|\mathbf{ZR}_{b,tl}(\tau, f)|$ and imaginary $|\mathbf{ZI}_{b,tl}(\tau, f)|$ parts are learned jointly using the GJDL algorithm and get the estimated signals $\widetilde{\mathbf{X}}\mathbf{1}_{b,tl}(\tau, f)$, $\widetilde{\mathbf{Y}}\mathbf{1}_{b,tl}(\tau, f)$ from the magnitude part and $\widetilde{\mathbf{X}}\mathbf{2}_{b,tl}(\tau, f)$, $\widetilde{\mathbf{Y}}\mathbf{2}_{b,tl}(\tau, f)$ from the real and imaginary parts by using LARC.

Let we get $\widetilde{\mathbf{X}}_{b,tl}(\tau, f)$ and $\widetilde{\mathbf{Y}}_{b,tl}(\tau, f)$ that are the estimated complex speech signals from $\widetilde{\mathbf{X}}\mathbf{1}_{b,tl}(\tau, f)$, $\widetilde{\mathbf{Y}}\mathbf{1}_{b,tl}(\tau, f)$ and $\widetilde{\mathbf{X}}\mathbf{2}_{b,tl}(\tau, f)$, $\widetilde{\mathbf{Y}}\mathbf{2}_{b,tl}(\tau, f)$ respectively by applying the Gini index. At last, the expected first and second source speech signals are calculated via the following equations.

$$\tilde{x}(t) = IDTCWT \left(ISTFT \left(\tilde{X}_{b,u}(\tau, f) \right) \right) \tag{5}$$

$$\tilde{y}(t) = IDTCWT \left(ISTFT \left(\tilde{Y}_{b,u}(\tau, f) \right) \right) \tag{6}$$

where $\tilde{x}(t)$ and $\tilde{y}(t)$ are the estimated first and second source signals, ISTFT and IDTCWT specify the inverse short term Fourier transform and inverse dual-tree complex wavelet transform, respectively.

3 Preliminaries

This section presents all relevant terms such as DTCWT, STFT, GDL, and GI that are linked to our proposed technique.

3.1 DTCWT

Kingsbury presents in [21] that the more computationally productive strategy refers to the DTCWT having different valuable properties, such as approximate shift-invariance, perfect reconstruction, and limited redundancy. The DTCWT [14] splits the signal into two trees; the first tree delivers the real part of the transform, while the second tree offers the imaginary part. Both trees have a low pass filter that provides the approximate coefficient and a high pass filter that delivers a detail coefficient. The complex-valued scaling functions and wavelets calculated from the two trees are roughly analytic.

In the first level DTCWT decomposition, both trees have one approximation coefficient and one detail coefficient. For the upper tree, the approximation coefficient is $(x_{1,1}^1)$ and the detail coefficient is $(x_{2,1}^1)$, here, the DTCWT decomposition level (dl) is presented by superscript, and the first and second subscripts indicate the subband index and tree-level (tl), individually. At that point, all subband signals are downsampled. For the second level decomposition, the filters are used to pass through the approximation coefficients only, and the subband signals are produced and so on. Two levels of the DTCWT decomposition are shown in Fig. 1a and b represent IDTCWT.

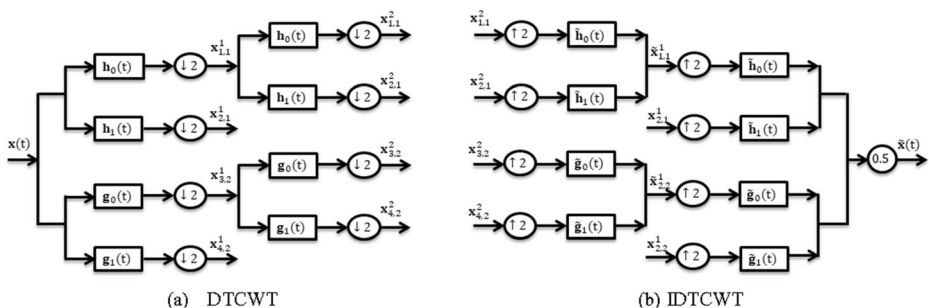


Fig. 1 The two-level filter bank implementation of a the DTCWT and b the IDTCWT

3.2 STFT

STFT is a dominant time-frequency analysis tool for audio signal processing [1]. It illustrates an especially useful class of time-frequency distributions which indicate complex amplitude versus time and frequency for any signal. In practice, the data to be transformed is divided into shorter segments of equal length. Each section is Fourier transformed separately, and the complex outcome is complementary to a matrix. It can be expressed as follows.

$$STFT \{x(t)\} = X(\tau, f) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-ift} dt \tag{7}$$

Here, $w(\tau)$ denotes window function, restrained around zero, and $x(t)$ is the signal to be transformed. $X(\tau, f)$ is basically the Fourier Transform of $x(t)w(t - \tau)$ that signifies the phase and magnitude of the signal over time and frequency. The time index τ is shifted windows, and f identifies the frequency.

3.3 GDL

The GDL approach is addressed in [33]. During the training process, the mix speech matrix $Z \in \mathbb{R}^F \times T$ is distributed approximately into two matrices; one is dictionary matrix $D \in \mathbb{R}^F \times R$ and another is coefficient matrix $C \in \mathbb{R}^R \times T$, and the number of dictionary atoms is symbolized by R . The sparse representation error of the speech signals x and y over the speech signals dictionary D_x , and D_y respectively can be reduced using Eq. (8) and Eq. (9), as follows:

$$\min_{D_x, C_x} \|X - D_x C_x\|_F^2 \quad \text{s.t. } \|c_{x,k}\|_1 \leq q_x, \forall k, \tag{8}$$

$$\min_{D_y, C_y} \|Y - D_y C_y\|_F^2 \quad \text{s.t. } \|c_{y,k}\|_1 \leq q_y, \forall k, \tag{9}$$

where $\|\cdot\|_F$ indicates the Frobenius norm, and $\|\cdot\|_1$ indicates the l_1 norm. The k^{th} column of the sparse coding matrix C_x and C_y , are presented by $c_{x,k}$ and $c_{y,k}$ respectively. q_x and q_y are the sparsity constraint for speech signals x and y , respectively. In [33], the LARC scheme is developed for sparse coding to solve the cost function presented in Eq. (8) and Eq. (9). K-SVD scheme is estimated for dictionary update, and D_x and D_y are obtained. The mixed signal is sparsely exemplified over the composite dictionary as follows

$$Z^{\text{test}} = D \times E = [D_x, D_y] \times \begin{bmatrix} E_x \\ E_y \end{bmatrix} \tag{10}$$

where E_x and E_y indicate the sparse coding matrix during the testing stage equivalent to D_x and D_y . Finally, the estimated speeches can be obtained in the following way.

$$\hat{X} = D_x \times E_x \tag{11}$$

$$\hat{Y} = D_y \times E_y \tag{12}$$

3.4 GI

The GI, introduced in 1921, is utilized to process the imbalance or sparseness of wealth or speech distribution. It is the principal measure that fulfills all the attractive standards for sparsity [10]. The GI is twice the area between the Lorenz curve and the 45-degree line.

Given data $\mathbf{z}(t) = [z(1), z(2), \dots, z(T)]$, the Lorenz curve initially defined in [24], is the function with support $(0, 1)$, which is piecewise linear with $T + 1$ points defined,

$$\mathbf{L}(g/T) = \sum_{j=1}^g [z_{(j)} / \sum_{k=1}^T z_k], \quad g = 0, \dots, T \quad (13)$$

where $z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(T)}$ denotes T values orderly from lowest to biggest. α indicates GI as the following equation.

$$\alpha = 1 - \frac{1}{T} \sum_{m=1}^T \left(\mathbf{L}\left(\frac{m-1}{T}\right) + \mathbf{L}\left(\frac{m}{T}\right) \right). \quad (14)$$

4 Proposed SS algorithm

In this section, we describe the newly proposed SS algorithm and subtleties connected to the substances of this algorithm. Most speech separation systems work on the STFT of the speech signal considering only the magnitude spectrum. Generally, the STFT transforms the time domain input signal by taking the small segments or frames of that signal and deliberates each subdivision to be stationary. But the subdivision may not be more stationary because we can't surely know what frequency exists at what time occurrence. In our proposed algorithm, firstly, we use DTCWT that divides the input signal into subbands where high and low-frequency components are separated. For a particular DTCWT level decomposition, we have represented all subbands by $\mathbf{x}_{b, tl}$, where $b = dl + 1$, and $tl = 2$ because it has 2-trees. For one level DTCWT decomposition, the total number of subbands is $4 (2 \times 2)$, and two levels DTCWT decomposition the total number of subbands is $6 (3 \times 2)$ and so on. In our proposed technique, we use the first-level decomposition, in which the time-domain signal is decomposed into four subband signals. For illustration, the DTCWT breaks down the source signal $x(t)$ into subband signals, denoted by $\mathbf{x}_{b, tl}$. For the first level decomposition, the subbands are $\mathbf{x}_{1, 1}$, $\mathbf{x}_{1, 2}$, $\mathbf{x}_{2, 1}$, and $\mathbf{x}_{2, 2}$, as clarified in the DTCWT part of Section 3. Then STFT is applied to each more stationary subband signal that comes from DTCWT decomposition. STFT gives superior transforms for more stationary signals. After applying DTCWT and STFT successively, the generative joint dictionary learning (GJDL) algorithm is used to jointly learn the magnitude, the absolute value of the real and imaginary parts of the signal. The LARC algorithm catches the required coefficients using such dictionaries. The initial signals are evaluated in two ways in our proposed method. The first is an estimated signal that only considers the magnitude component, whereas the second is an estimated signal that considers both the real and imaginary components. Figure 2 illustrates the comprehensive framework of the SS algorithm mentioned in this paper. We use these dual transformations in both the testing and training stages, and the training and testing phases are detailed separately in the following subsections.

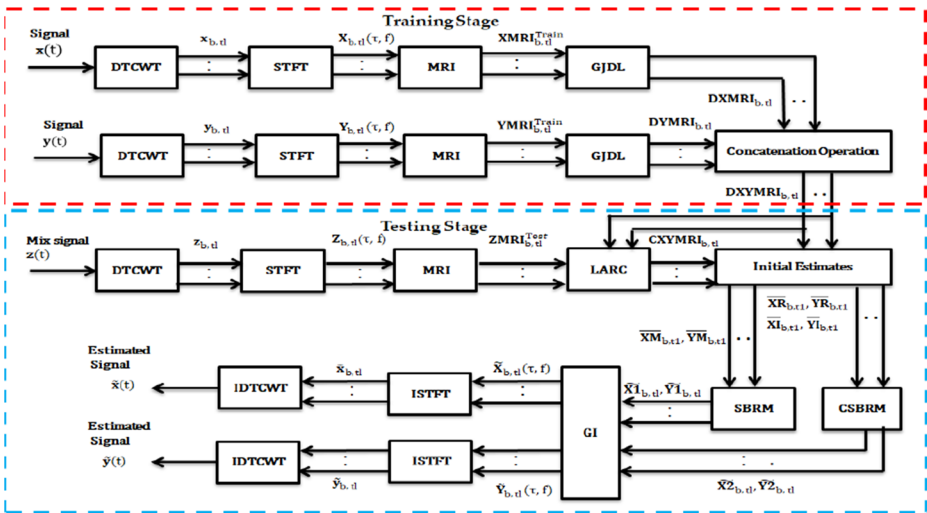


Fig. 2 Shows a block diagram of the proposed speech separation system including the training and testing stage

4.1 Training stage

In the training stage, we consider two individual speech source generating signals $x(t)$ and $y(t)$. DTCWT is utilized to get the subband source signals $x_{b,tl}$ and $y_{b,tl}$ from the signals $x(t)$ and $y(t)$. The STFT is applied to each subband source signal and found the complex spectrums $X_{b,tl}(\tau, f)$ and $Y_{b,tl}(\tau, f)$, where τ and f specify the time and frequency bin indexes, respectively. At this point, we obtain three parts the magnitude part $XM_{b,tl}(\tau, f)$, the real part $XR_{b,tl}(\tau, f)$ and the imaginary parts $XI_{b,tl}(\tau, f)$ from $X_{b,tl}(\tau, f)$, and apply similar operation for $Y_{b,tl}(\tau, f)$. We take the absolute of the real and imaginary parts and concatenate it with a magnitude part as follows.

$$XMRI_{b,tl}^{Train} = \begin{bmatrix} |XM_{b,tl}(\tau, f)| \\ |XR_{b,tl}(\tau, f)| \\ |XI_{b,tl}(\tau, f)| \end{bmatrix} \tag{15}$$

$$YMRI_{b,tl}^{Train} = \begin{bmatrix} |YM_{b,tl}(\tau, f)| \\ |YR_{b,tl}(\tau, f)| \\ |YI_{b,tl}(\tau, f)| \end{bmatrix} \tag{16}$$

The GJDL is applied to train the dictionaries $DXMRI_{b,tl}$ and $DYMRI_{b,tl}$ for all of the three components by using Eqs. (8) and (9), based on Eqs. (15) and (16). Using the LARC algorithm for sparse coding and the approximate K-SVD algorithm for dictionary update [33], the cost function Eqs. (8) and (9) have been solved and $DXMRI_{b,tl}$ and $DYMRI_{b,tl}$ are acquired as follows.

$$DXMRI_{b,tl} = GJDL(XMRI_{b,tl}^{Train}) \tag{17}$$

$$\mathbf{DYMRI}_{b,tl} = \mathbf{GJDL} \left(\mathbf{YMRI}_{b,tl}^{\text{Train}} \right) \quad (18)$$

At this point, we concatenate all the dictionaries and obtain the concatenated dictionaries like $\mathbf{DXYMRI}_{b,tl}$, in eq. (19) and it is forwarded to the testing phase.

$$\mathbf{DXYMRI}_{b,tl} = \left[\mathbf{DXMRI}_{b,tl} \quad \mathbf{DYMRI}_{b,tl} \right] \quad (19)$$

4.2 Testing stage

The mixed speech signal $\mathbf{z}(t)$ is decomposed by applying DTCWT and produced a set of subband signals $\mathbf{z}_{b,tl}$. The STFT is applied to every subband of mixed signal and obtained complex spectrum $\mathbf{Z}_{b,tl}(\tau, f)$. At this point, we take the magnitude part $|\mathbf{ZM}_{b,tl}(\tau, f)|$, phase $\mathbf{ZP}_{b,tl}$, the real $\mathbf{ZR}_{b,tl}(\tau, f)$ and the imaginary $\mathbf{ZI}_{b,tl}(\tau, f)$ parts from complex spectrums $\mathbf{Z}_{b,tl}(\tau, f)$ and preserve sign values of real and imaginary parts. We take the absolute of the real and imaginary parts and concatenate it with a magnitude part as follows.

$$\mathbf{ZMRI}_{b,tl}^{\text{Test}} = \begin{bmatrix} |\mathbf{ZM}_{b,tl}(\tau, f)| \\ |\mathbf{ZR}_{b,tl}(\tau, f)| \\ |\mathbf{ZI}_{b,tl}(\tau, f)| \end{bmatrix} \quad (20)$$

We have the testing mixture signal with three parts $\mathbf{ZMRI}_{b,tl}^{\text{Test}}$ and the concatenated dictionaries $\mathbf{DXYMRI}_{b,tl}$, we can obtain the sparse coding $\mathbf{CXYMRI}_{b,tl}$ by using eq. (10) as follows.

$$\mathbf{CXYMRI}_{b,tl} = \mathbf{LARC} \left(\mathbf{ZMRI}_{b,tl}^{\text{Test}}, \mathbf{DXYMRI}_{b,tl} \right) \quad (21)$$

The initially estimated magnitude, real and imaginary components $\overline{\mathbf{XM}}_{b,tl}$, $\overline{\mathbf{XR}}_{b,tl}$, and $\overline{\mathbf{XI}}_{b,tl}$, respectively for one source signal and $\overline{\mathbf{YM}}_{b,tl}$, $\overline{\mathbf{YR}}_{b,tl}$, and $\overline{\mathbf{YI}}_{b,tl}$ for another source signal are acquired using the corresponding speech signals dictionaries and sparse coding, which are getting from $\mathbf{DXYMRI}_{b,tl}$ and $\mathbf{CXYMRI}_{b,tl}$ as follows.

$$\overline{\mathbf{XM}}_{b,tl} = \mathbf{DXM}_{b,tl} \mathbf{CX}_{b,tl}, \quad (22)$$

$$\overline{\mathbf{XR}}_{b,tl} = \mathbf{DXR}_{b,tl} \mathbf{CX}_{b,tl}, \quad (23)$$

$$\overline{\mathbf{XI}}_{b,tl} = \mathbf{DXI}_{b,tl} \mathbf{CX}_{b,tl}, \quad (24)$$

$$\overline{\mathbf{YM}}_{b,tl} = \mathbf{DYM}_{b,tl} \mathbf{CY}_{b,tl}, \quad (25)$$

$$\overline{\mathbf{YR}}_{b,tl} = \mathbf{DYR}_{b,tl} \mathbf{CY}_{b,tl}, \quad (26)$$

$$\widetilde{\mathbf{Y}}_{b,tl} = \mathbf{D}\mathbf{Y}_{b,tl}\mathbf{C}\mathbf{Y}_{b,tl}, \tag{27}$$

The addition of the initial estimate $\widetilde{\mathbf{X}}\mathbf{M}_{b,tl}$ and $\widetilde{\mathbf{Y}}\mathbf{M}_{b,tl}$ may not be equivalent to the mixed signal magnitude spectrum $\mathbf{Z}\mathbf{M}_{b,tl}$. To make it error-free, we compute the subband ratio mask (SBRM) using the following Eq. (28) and Eq. (29) as follows:

$$\widetilde{\mathbf{X}}\mathbf{I}_{b,tl} = \frac{\left(\widetilde{\mathbf{X}}\mathbf{M}_{b,tl}\right)^2}{\left(\widetilde{\mathbf{X}}\mathbf{M}_{b,tl}\right)^2 + \left(\widetilde{\mathbf{Y}}\mathbf{M}_{b,tl}\right)^2} \times \mathbf{Z}\mathbf{M}_{b,tl}, \tag{28}$$

$$\widetilde{\mathbf{Y}}\mathbf{I}_{b,tl} = \frac{\left(\widetilde{\mathbf{Y}}\mathbf{M}_{b,tl}\right)^2}{\left(\widetilde{\mathbf{X}}\mathbf{M}_{b,tl}\right)^2 + \left(\widetilde{\mathbf{Y}}\mathbf{M}_{b,tl}\right)^2} \times \mathbf{Z}\mathbf{M}_{b,tl}, \tag{29}$$

Now, we apply the phase spectrum $\mathbf{Z}\mathbf{P}_{b,tl}$ with the estimated source signals magnitude spectrum $\widetilde{\mathbf{X}}\mathbf{I}_{b,tl}$ and $\widetilde{\mathbf{Y}}\mathbf{I}_{b,tl}$ to acquire the reformed complex spectrum $\widetilde{\mathbf{X}}\mathbf{I}_{b,tl}(\tau, f)$ and $\widetilde{\mathbf{Y}}\mathbf{I}_{b,tl}(\tau, f)$ using the Eq. (30) and Eq. (31) as follows:

$$\widetilde{\mathbf{X}}\mathbf{I}_{b,tl}(\tau, f) = \widetilde{\mathbf{X}}\mathbf{I}_{b,tl} e^{i\mathbf{Z}\mathbf{P}_{b,tl}}, \tag{30}$$

$$\widetilde{\mathbf{Y}}\mathbf{I}_{b,tl}(\tau, f) = \widetilde{\mathbf{Y}}\mathbf{I}_{b,tl} e^{i\mathbf{Z}\mathbf{P}_{b,tl}}, \tag{31}$$

Now, we use the sign preserved previously and multiply the sign with real and imaginary estimates of the signal. Then, the real and imaginary parts are joined to form the complex spectrum of the speech signals as follows.

$$\widetilde{\mathbf{X}}\mathbf{2}_{b,tl} = \widetilde{\mathbf{X}}\mathbf{R}_{b,tl} + i\widetilde{\mathbf{X}}\mathbf{I}_{b,tl}, \tag{32}$$

$$\widetilde{\mathbf{Y}}\mathbf{2}_{b,tl} = \widetilde{\mathbf{Y}}\mathbf{R}_{b,tl} + i\widetilde{\mathbf{Y}}\mathbf{I}_{b,tl}, \tag{33}$$

To make the estimated signal $\widetilde{\mathbf{X}}\mathbf{2}_{b,tl}$ and $\widetilde{\mathbf{Y}}\mathbf{2}_{b,tl}$ are more accurate, we calculate the complex subband ratio mask (CSBRM) using the following Eq. (34) and Eq. (35).

$$\widetilde{\mathbf{X}}\mathbf{2}_{b,tl}(\tau, f) = \frac{\left(\widetilde{\mathbf{X}}\mathbf{2}_{b,tl}\right)^2}{\left(\widetilde{\mathbf{X}}\mathbf{2}_{b,tl}\right)^2 + \left(\widetilde{\mathbf{Y}}\mathbf{2}_{b,tl}\right)^2} \times \mathbf{Z}_{b,tl}(\tau, f), \tag{34}$$

$$\widetilde{\mathbf{Y}}\mathbf{2}_{b,tl}(\tau, f) = \frac{\left(\widetilde{\mathbf{Y}}\mathbf{2}_{b,tl}\right)^2}{\left(\widetilde{\mathbf{X}}\mathbf{2}_{b,tl}\right)^2 + \left(\widetilde{\mathbf{Y}}\mathbf{2}_{b,tl}\right)^2} \times \mathbf{Z}_{b,tl}(\tau, f), \tag{35}$$

The accuracy of $\widetilde{\mathbf{X}}_{b,tl}(\tau, f)$ and $\widetilde{\mathbf{X}}_{2,b,tl}(\tau, f)$ are not similar due to the different estimation processes. The first is based on the signal's magnitude, while the second is based on the signal's real and imaginary components. $\widetilde{\mathbf{Y}}_{b,tl}(\tau, f)$ and $\widetilde{\mathbf{Y}}_{2,b,tl}(\tau, f)$ are also estimated in the same way. As these estimated signals have complementary effectiveness we use a weighting parameter $\alpha_{b,tl}$ which is found by using Eq. (14), and estimated signals can be calculated as follows:

$$\widetilde{\mathbf{X}}_{b,tl}(\tau, f) = (1 - \alpha_{b,tl})\widetilde{\mathbf{X}}_{1,b,tl}(\tau, f) + \alpha_{b,tl}\widetilde{\mathbf{X}}_{2,b,tl}(\tau, f), \quad (36)$$

$$\widetilde{\mathbf{Y}}_{b,tl}(\tau, f) = (1 - \alpha_{b,tl})\widetilde{\mathbf{Y}}_{1,b,tl}(\tau, f) + \alpha_{b,tl}\widetilde{\mathbf{Y}}_{2,b,tl}(\tau, f), \quad (37)$$

The ISTFT is used to transform the complex speech signals spectrum $\widetilde{\mathbf{X}}_{b,tl}(\tau, f)$ and $\widetilde{\mathbf{Y}}_{b,tl}(\tau, f)$ to the subband signals $\widetilde{\mathbf{x}}_{b,tl}$ and $\widetilde{\mathbf{y}}_{b,tl}$. Finally, the estimated source speech signals $\widetilde{\mathbf{x}}(t)$ and $\widetilde{\mathbf{y}}(t)$ is achieved by transforming the IDTCWT to the subband signals $\widetilde{\mathbf{x}}_{b,tl}$ and $\widetilde{\mathbf{y}}_{b,tl}$. The suggested algorithm for the training and testing stages are presented below in Table 2.

5 Evaluation and results

In this section, the proposed algorithm is analyzed through simulation experiments. First, provide an overview of the data and performance evaluation indicators that will be used to measure the efficiency of separated speech. We show the impact of joint learning regarding SDR, SIR, STOI, PESQ, HASPI, and HASQI scores at male-female separation. Then we explore the impact of GJDL over SNMF concerning STOI and PESQ scores at the same and opposite gender cases. Finally, we compare our algorithm with the current mainstream single-channel SS algorithm and use the experimental results to confirm the lead of the proposed strategy. The comparison algorithms are STFT-SNMF [45], DWT-STFT-SNMF [42], ADDL [2], SWT-SNMF [11], DTCWT-SNMF [12], CJD [35], OJDL [36], and DTCWT-STFT-SNMF [8].

5.1 Data sets and performance evaluation indicators

In this simulation, we collect the speech signals (including different male speech and female speech) from the GRID audio-visual corpuses [3], which are used as the training and testing data. There are 34 speakers (18 male, 16 female), and each speaker speaks 1000 utterances. In case of selecting each speakers' utterances, we randomly take 500 utterances for training purposes and 200 utterances for testing. In this simulation, we use two types of speech signal data grouping; one is used for same-gender (male-male or female-female) speech separation and another for opposite gender (male-female) speech separation. For same-gender speech separation, eight same-gender speakers' utterances are exploited to form one experimental group, and different eight same-gender speakers' utterances are used to build another experimental group. For opposite-gender speech separation, we choose sixteen male speakers for one experimental group and sixteen females for another experimental group. The length of the training signal is about 60 s, and that for the test is about 10 s. The sampling rate of a speech signal is 8000 Hz, and the signal is transformed into the time-frequency domain by using 512-point STFT.

Table 2 Algorithm for the training and testing stages of the proposed technique

<p>Training Algorithm: Input: Training sets $x(t)$ and $y(t)$, decomposition level (dl), and tree-level (tl). Output: $\text{DXMRI}_{p,q}$. Step 1: Set $p=1$ to tl Step 2: Set $q=1$ to b, where $b = dl + 1$. Step 3: Determine the wavelet coefficients via DTCWT. $x_{p,q} = \text{DTCWT}(x(t))$ and $y_{p,q} = \text{DTCWT}(y(t))$. Step 4: Take a complex spectrum, applying STFT. $X_{p,q} = \text{STFT}(x_{p,q})$ and $Y_{p,q} = \text{STFT}(y_{p,q})$. Step 5: Concatenate the magnitude spectrum and absolute value of the real and imaginary components to form $\text{XMRI}_{p,q}$ and $\text{YMRI}_{p,q}$. Step 6: Determine the basis of dictionaries. $\text{DXMRI}_{p,q} = \text{GJDL}(\text{XMRI}_{p,q})$ and $\text{DYMRI}_{p,q} = \text{GJDL}(\text{YMRI}_{p,q})$. Step 7: Concatenate these basis dictionaries like as $\text{DXMRI}_{p,q} = [\text{DXMRI}_{p,q} \text{DYMRI}_{p,q}]$. Step 8: $q=q+1$, go to step 2. Step 9: $p=p+1$, go to step 1. Testing Algorithm: Input: Mixed signal $z(t)$, concatenated basis dictionaries $\text{DXMRI}_{p,q}$ learned in the training stage, decomposition level (dl), and tree-level (tl). Output: Estimated separate signals $\hat{x}(t)$ and $\hat{y}(t)$. Step 1: Set $p=1$ to tl Step 2: Set $q=1$ to b, where $b = dl + 1$. Step 3: Calculate the wavelet coefficients, $z_{p,q} = \text{DTCWT}(z(t))$. Step 4: Obtain a complex spectrum, $Z_{p,q} = \text{STFT}(z_{p,q})$. Step 5: Compute magnitude, phase, absolute value, and sign of real and imaginary components from $Z_{p,q}$. Step 6: Concatenate the magnitude and absolute value of the real and imaginary components, $\text{ZMRI}_{p,q}$. Step 7: Obtain the coefficient matrices, according to Eq. (21). $\text{DZMRI}_{p,q} = \text{LARC}(\text{ZMRI}_{p,q}, \text{DXMRI}_{p,q})$ Step 8: Estimate the initial magnitude, real and imaginary components by using Eq. (22) to Eq. (27). Step 9: Compute the subband binary ratio masks considering the only magnitude and obtain $\overline{\text{X1}}_{p,q}$ and $\overline{\text{Y1}}_{p,q}$. Step 10: Determine the first estimated complex spectrogram $\overline{\text{X1}}_{p,q}(\tau, f)$ and $\overline{\text{Y1}}_{p,q}(\tau, f)$. Step 11: Multiply sign with real and imaginary components and make complex to obtain $\overline{\text{X2}}_{p,q}$ and $\overline{\text{Y2}}_{p,q}$. Step 12: Calculate the complex subband binary ratio masks and get the second estimated complex spectrogram signals $\overline{\text{X2}}_{p,q}(\tau, f)$ and $\overline{\text{Y2}}_{p,q}(\tau, f)$. Step 13: Apply the Gini index. $\tilde{\text{X}}_{p,q}(\tau, f) = \text{GI}(\overline{\text{X1}}_{p,q}(\tau, f), \overline{\text{X2}}_{p,q}(\tau, f))$ and $\tilde{\text{Y}}_{p,q}(\tau, f) = \text{GI}(\overline{\text{Y1}}_{p,q}(\tau, f), \overline{\text{Y2}}_{p,q}(\tau, f))$. Step 14: Compute the modified subband signals. $\tilde{x}_{p,q} = \text{ISTFT}(\tilde{\text{X}}_{p,q}(\tau, f))$ and $\tilde{y}_{p,q} = \text{ISTFT}(\tilde{\text{Y}}_{p,q}(\tau, f))$. Step 15: $q=q+1$, go to step 2. Step 16: $p=p+1$, go to step 1. Step 17: Obtain the estimated source signals $\hat{x}(t)$ and $\hat{y}(t)$ by applying the IDTCWT.</p>

In this paper, the following six indicators are used to measure the performance of SS: HASQI [18], HASPI [19], PESQ [19], STOI [40], SDR [43], and SIR [43] metrics.

The SDR [43] value approximates the overall speech quality, and it is the proportion of the intensity of the input signal to the intensity of the difference among input and reformed signals. The higher SDR scores regulate the recovered performance.

$$\text{SDR} = 10 \log_{10} \frac{\|x_{\text{target}}\|_1^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|_1^2} \tag{38}$$

where x_{target} , e_{interf} , e_{noise} , and, e_{artif} are the targeted source, the interference error, the perturbation noise and the artifacts error, respectively.

In addition to SDR, SIR [43] reports errors produced by failures to remove the interfering signal during the source separation technique. A higher value of SIR relates to higher separation quality.

$$\text{SIR} = 10 \log_{10} \frac{\|\mathbf{x}_{\text{target}}\|_1^2}{\|\mathbf{e}_{\text{interf}}\|_1^2} \quad (39)$$

The PESQ is nominated for the objective quality assessment and is commonly used to measure speech signals' quality superiority. It deals with scores ranging from -0.50 to 4.50 , where the higher scores lead to more outstanding voice quality. PESQ measures the combination of only two parameters – one symmetric disturbance (d_{SYM}) and one asymmetric disturbance (d_{ASYM}), provides a good balance between prediction accuracy and the ability to simplify, described in [19].

$$\text{PESQ} = 4.5 - 0.1 d_{\text{SYM}} - 0.0309 d_{\text{ASYM}} \quad (40)$$

STOI [40] is a state-of-the-art speech intelligibility indicator and deals with the correlation coefficient among the clean speech temporal envelopes and the separated speech in short-time regions. It conveys the scores somewhere in the range of 0 to 1, where the higher STOI value signifies better intelligibility. We measure STOI, in light of a correlation coefficient between the transient envelopes of the perfect and estimated speech, in a short time frame overlapping fragments. It is a function of the clean and corrupted speech, represented by \mathbf{x} and $\tilde{\mathbf{x}}$, respectively.

$$\text{STOI} = \text{Avg} \left(\frac{(\mathbf{x} - \mu_{\mathbf{x}})^T (\tilde{\mathbf{x}} - \mu_{\tilde{\mathbf{x}}})}{\|\mathbf{x} - \mu_{\mathbf{x}}\|_1 \|\tilde{\mathbf{x}} - \mu_{\tilde{\mathbf{x}}}\|_1} \right) \quad (41)$$

The HASPI [19] depends on a model of the hear-able fringe that incorporates changes because of hearing loss. The file looks at the envelope and fleeting fine construction yields of the hear-able model for a reference signal to the yields of the model for the signal below test. It ranges from 0 to 1, and advanced scores relay to better sound intelligibility.

The HASPI intelligibility index is specified by:

$$p = -9.047 + 14.817c + 0.0a_{\text{Low}} + 0.0a_{\text{Mid}} + 4.616a_{\text{High}} \quad (42)$$

$$\text{HASPI} = \frac{1}{1 + e^{-p}}$$

where c is cepstral correlation and a_{Low} , a_{Mid} , and a_{High} are the low-level auditory coherence value, mid-level value, and high-level value, respectively.

The HASQI [18] is a model-based objective measure of quality created regarding portable hearing assistants for ordinary hearing and hearing-impaired listeners. HASQI is the product of two autonomous indices. The first Q_{Nonlin} , detentions the properties of noise and nonlinear distortion, and the second, Q_{Linear} , detentions the properties of linear filtering and spectral

changes by focusing on contrasts in the long-term average spectra. It ranges from 0 to 1, and advanced scores relay to better sound quality.

$$\begin{aligned}
 Q_{\text{Linear}} &= 1 - 0.579\sigma_1 - 0.421\sigma_2 \\
 Q_{\text{Nonlin}} &= c^3 \\
 \text{HASQI} &= Q_{\text{Nonlin}} \times Q_{\text{Linear}}
 \end{aligned}
 \tag{43}$$

where c , σ_1 , σ_2 are cepstrum correlation, standard deviation of the spectral difference, and the standard deviation of the slope difference, respectively.

5.2 Impact of joint learning

The speech signal is short-term stationary and sparse in nature. Most of the speech separation methods used STFT to transform the signal into the time-frequency domain, which makes the complex spectrum of that signal. Some approaches have been proposed considering only the magnitude part while overlooking the real and imaginary part of a complex spectrum. Here we compare among the methods deliberate only the magnitude part, only the real and imaginary part and the magnitude, real, and imaginary part jointly. Figure 3 shows that if we use joint learning (magnitude, real and imaginary part of the complex matrix), it beats the SDR, SIR, HASQI, and STOI scores in male-female separation to the individual (magnitude or real and imaginary). That’s why in the proposed approach, we consider the magnitude, the real and imaginary parts together, which upgrades the separation performance.

5.3 Effect of GJDL over SNMF

We are exploring the impact of GJDL over SNMF concerning STOI and PESQ scores at the same and opposite gender cases. Figure 4 reveals the GJDL’s impact on SNMF. For all considering cases, PESQ and STOI scores are produced and averaged. The DTCWT-STFTMRI-SNMF and DTCWT-STFTMRI-GJDL methods use the SNMF and GJDL, respectively. It appears to be shown that both the speech’s PESQ and STOI values are improved at the same and opposite gender cases, explaining that to some degree, the DTCWT-STFTMRI-GJDL takes care of the speech signals distortion issues subsequently the SS processing.

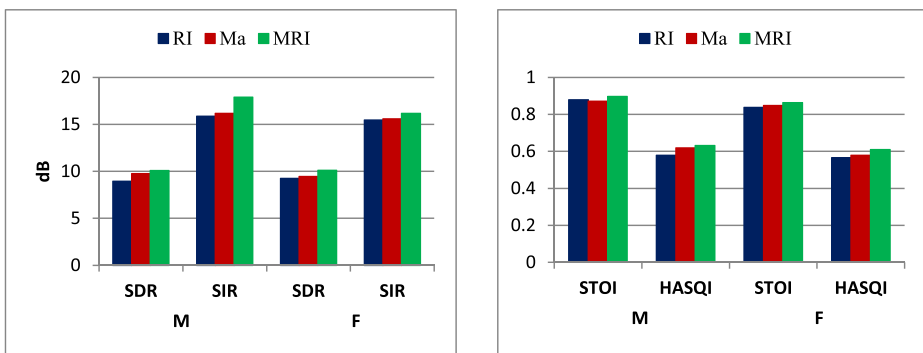


Fig. 3 Effect of joint learning at opposite gender cases (Ma: magnitude, RI: real and imaginary and MRI: magnitude, real and imaginary)

5.4 Overall performance comparison of algorithms

In Fig. 5, we show that the SDR and SIR of the proposed model gain considerably better results than other existing models, namely STFT-SNMF, DWT-STFT-SNMF, ADDL, SWT-SNMF, DTCWT-SNMF, CJD, and OJDL. For all cases of separation, the SDR values of the proposed model are higher than the existing models. Our proposed method increases SDR scores by 37.47%, 39.17% for M1 and M2, 21.29% and 18.20% for F1 and F2, and 27.73% and 27.20% for M and F, respectively than the existing method OJDL. From the figure, we can also realize that the SIR values of estimated signals are better than existing models. DTCWT and STFT are used consecutively for the dual transformation of the signal that delivers a more flexible basic framework for the improvement of feature modules that's why it gives better performance.

Fig. 6 presents the comparative performance analysis in terms of STOI and PESQ using the proposed method and other existing methods. STOI is improved from 0.746 to 0.819 for M1, 0.768 to 0.825 for M2, 0.785 to 0.800 for F1, 0.787 to 0.799 for F2, 0.793 to 0.896 for M and 0.778 to 0.863 for F using the proposed models over OJDL method. From the figure, we can also realize that the PESQ score of expected signals is better than the existing models. The STOI and PESQ scores in three different cases are shown that the suggested technique beats the other eight methods, i.e., the suggested approach deals with the maximum quality of speech separation comparative to the different seven schemes. In the case of considering only the magnitude part, the phase information is not enriched. But, if the complex domain training targets were exploited, then the phase information can be considered. As a result, we have used both magnitude, real, and imaginary components in our technique, which increased separation performance.

Tables 3 and 4 delineate the HASQI and HASPI results of different techniques, including STFT-SNMF, DWT-STFT-SNMF, ADDL, SWT-SNMF, DTCWT-SNMF, CJD, OJDL, DTCWT-STFT-SNMF, and DTCWT-STFTMRI-GJDL for the same gender and opposite gender speech separation. From Table 3, we can perceive that DTCWT-STFTMRI-GJDL earnings advanced HASPI values for all separation cases. It can also be seen that the HASQI results of DTCWT-STFTMRI-GJDL achieve progressive value to the other nine methods for all separation cases.

In order to have more performance evaluation about the proposed method, the spectrograms of the speech separation algorithms are examined. The separation results of the different approaches are displayed in Fig. 7, where the original female and male speech spectrograms are presented in Fig. 7a and b, respectively. The projected female and male speech

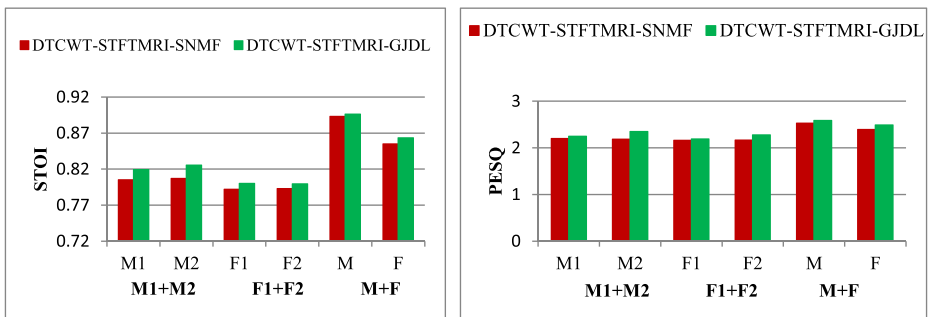
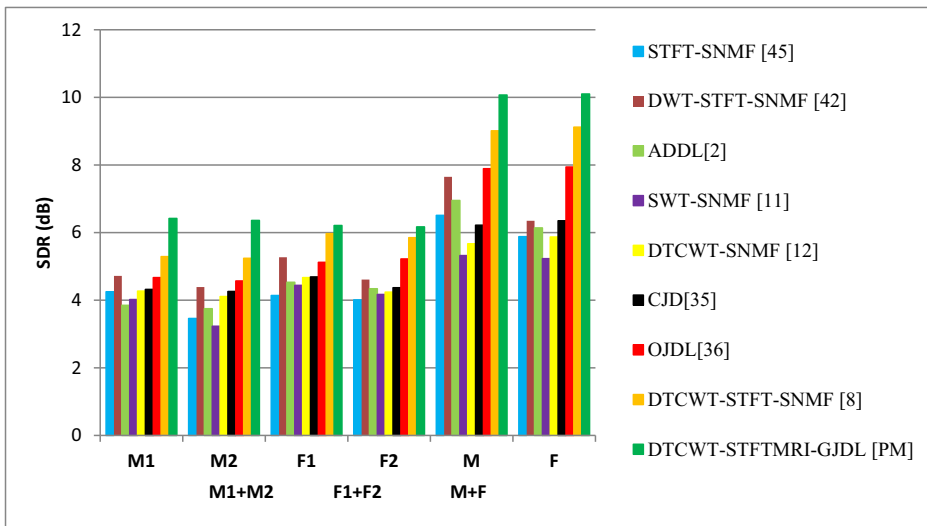
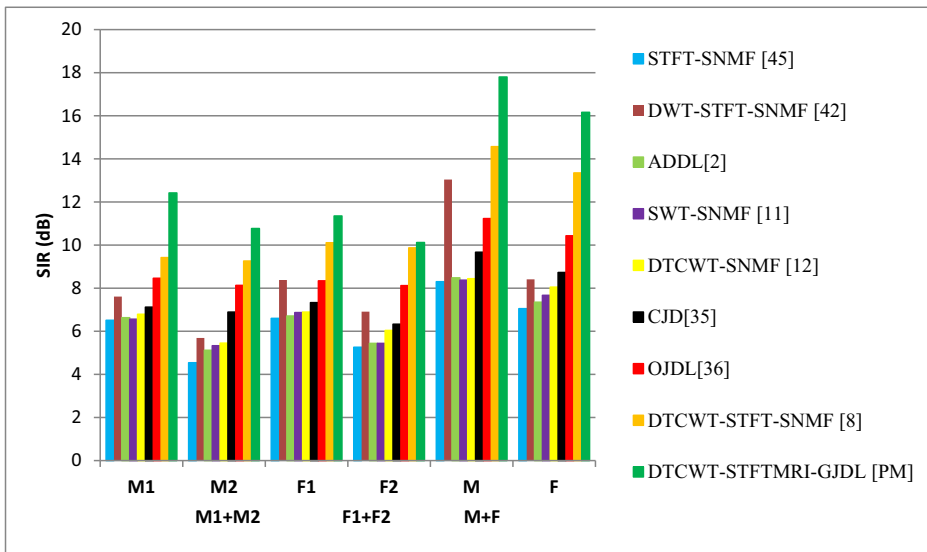


Fig. 4 Effect of GJDL over SNMF method concerning PESQ and STOI at the same and opposite gender cases



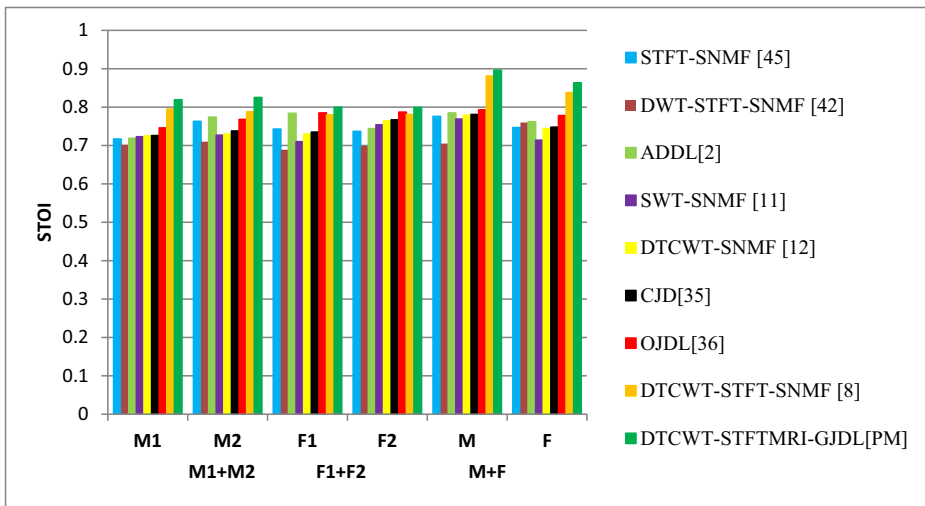
(a)



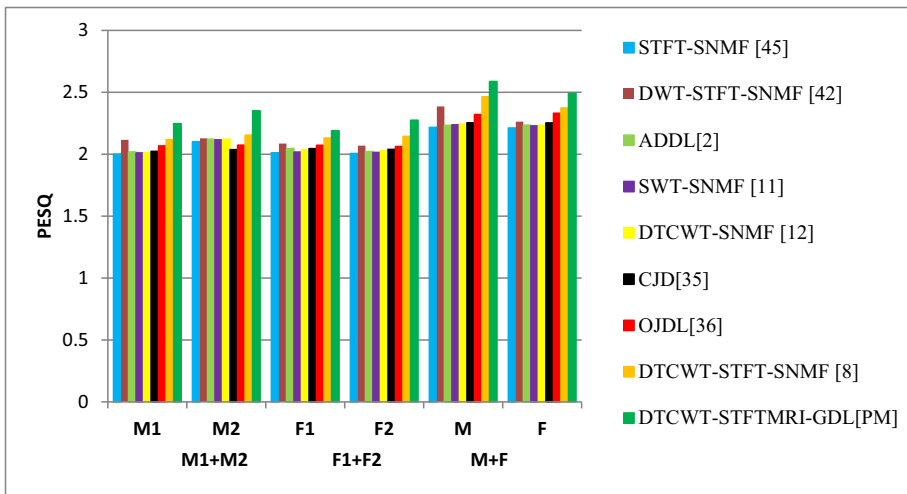
(b)

Fig. 5 Comparison of the separation performance of the nine methods in terms of **a** SDR and **b** SIR for the same and opposite gender cases

spectrograms are presented in Fig. 7c, d, e, f, g and h for DWT-STFT-SNMF, SWT-SNMF, and proposed model, respectively. From the figure, we see that the excellence of separated speech is poor in the DWT-STFT-SNMF method due to the entire elimination of the high-frequency component and estimated male and female speech by SWT-SNMF method supplement more undesirable vocal. The proposed method improves male and female speech roughly



(a)



(b)

Fig. 6 Comparison of the separation performance of the nine methods in terms of **a** STOI and **b** PESQ for the same and opposite gender cases

similar to original female and male speech. Also, it is realized that other methods have extra vocal distortion than our mentioned algorithm.

Finally, to further confirm the centralities of improvements, for a mixed speech separation investigation, we have used the TIMIT database [6]. We have investigated 24 speakers (12 male and 12 female speakers) were picked from the TIMIT database. Each speaker utters the ten sentences that outcomes total of 240 sentences. Out of 10 sentences of different speakers, the first eight sentences are selected for training, and the remaining are used for testing. To investigate the performance of our proposed scheme, we consider SDR, SIR, STOI, and PESQ

Table 3 Performance comparison among different techniques concerning HASPI values for the same and opposite gender cases

Case method	M1+M2		F1+F2		M+F	
	M1	M2	F1	F2	M	F
STFT-SNMF [45]	0.9942	0.9956	0.9954	0.9956	0.9982	0.9987
DWT-STFT-SNMF [42]	0.9724	0.9573	0.9761	0.9576	0.9951	0.9915
ADDL [2]	0.9781	0.9881	0.9871	0.9787	0.9898	0.9885
SWT-SNMF [11]	0.9767	0.99	0.9845	0.9854	0.9939	0.7514
DTCWT-SNMF [12]	0.9785	0.9911	0.9857	0.9817	0.9917	0.9894
CJD [35]	0.9793	0.9944	0.9863	0.9845	0.9943	0.9869
OJDL [36]	0.9813	0.9947	0.9896	0.9885	0.9964	0.9878
DTCWT-STFT-SNMF [8]	0.9961	0.9967	0.9971	0.9964	0.9995	0.9878
DTCWT-STFTMRI-GJDL [PM]	0.9978	0.9986	0.9976	0.9973	0.9997	0.9995

Table 4 Performance comparison of various approaches in terms of HASQI values for the same and opposite gender cases

Case method	M1+M2		F1+F2		M+F	
	M1	M2	F1	F2	M	F
STFT-SNMF [45]	0.412	0.405	0.407	0.414	0.555	0.503
DWT-STFT-SNMF [42]	0.269	0.246	0.271	0.257	0.485	0.445
ADDL [2]	0.421	0.412	0.413	0.434	0.565	0.513
SWT-SNMF [11]	0.322	0.352	0.328	0.428	0.488	0.418
DTCWT-SNMF [12]	0.333	0.364	0.33	0.42	0.47	0.452
CJD [35]	0.346	0.373	0.349	0.421	0.483	0.461
OJDL [36]	0.389	0.398	0.401	0.429	0.521	0.493
DTCWT-STFT-SNMF [8]	0.439	0.398	0.401	0.429	0.521	0.493
DTCWT-STFTMRI-GJDL [PM]	0.479	0.500	0.465	0.463	0.631	0.610

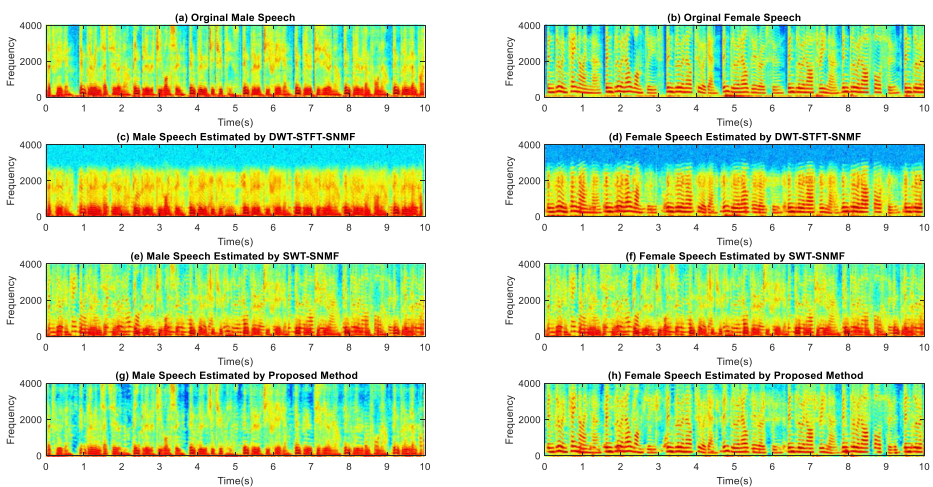


Fig. 7 Spectrogram of original male speech, original female speech, and recovered female speech, recovered male speech for DWT-STFT-SNMF, SWT-SNMF, and proposed model, where x-axis corresponds to the time in seconds, and the y-axis corresponds to the frequency in kHz

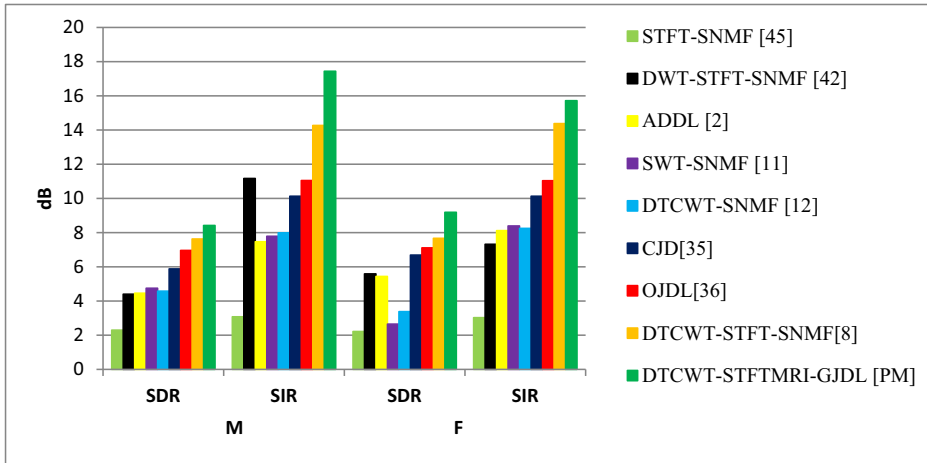


Fig. 8 Comparative performance evaluation of the existing and proposed model of SDR and SIR for the opposite gender case considering the TIMIT database

scores. From Fig. 8 and Table 5, one can without much stretch be seen that the proposed approach is achieved better performance than the existing eight techniques (STFT-SNMF, DWT-STFT-SNMF, ADDL, SWT-SNMF, DTCWT-SNMF, CJD, OJDL, and DTCWT-STFT-SNMF) depending on the SDR, SIR, STOI and PESQ scores at opposite gender separation.

6 Conclusion

We have developed a new framework of speech separation based on dual-domain transform in which GJDL is used for joint learning and GI for joint accuracy. The main emphasis is to learn the dictionary considering magnitude, real and imaginary parts jointly, in contrast to the traditional approach of learning considering only the magnitude part or only the complex domain. DTCWT and STFT are used serially for the dual transformation of the signal that offers a more flexible basic framework for upgrading feature segments. Then GJDL is used to

Table 5 Performance assessment of PESQ and STOI values of nine methods for the opposite gender case considering the TIMIT database

Method	PESQ		STOI	
	M	F	M	F
STFT-SNMF [45]	2.204	1.097	0.757	0.439
DWT-STFT-SNMF [42]	2.376	2.170	0.708	0.673
ADDL [2]	2.253	1.998	0.773	0.730
SWT-SNMF [11]	2.356	1.983	0.787	0.721
DTCWT-SNMF [12]	2.335	1.938	0.780	0.717
CJD [35]	2.324	2.012	0.789	0.719
OJDL [36]	2.375	2.141	0.805	0.730
DTCWT-STFT-SNMF [8]	2.509	2.192	0.805	0.730
DTCWT-STFTMRI-GJDL [PM]	2.666	2.349	0.855	0.805

jointly learn the magnitude, the absolute value of the real and imaginary parts of the signal. The LARC algorithm captures the required coefficients using such dictionaries. We initially estimate the signals in two ways, one by considering only the magnitude part and another in view of real and imaginary components. At last, the GI finds better accuracy analyzing the corresponding subband of the two different sets of estimated signals and achieves the final estimated speech signals.

The DTCWT separates the high and low-frequency components of the time domain signal, whereas the STFT accurately investigates the time-frequency components. We also deal with the signal's magnitude as well as its real and imaginary portions. As a result, this algorithm entirely uses all of the information contained in the waveforms of the signals. The relevant experimental results reveal that our approach outperforms traditional methods when measured using several evaluation metrics such as SDR, SIR, HASQI, HASPI, PESQ, and STOI. We use limited features to train GJDL but to get better performance need to consider more features. If we consider more features, the time complexity will increase for both training and testing stages. We plan to investigate alternative training and testing algorithms using deep neural networks in the future and expand it on multisource/multichannel processing, which is a very relevant and interesting path.

Acknowledgments This research was supported by the National Natural Science Foundation of China (no. 61671418).

Data availability The datasets created or analyzed during the present study are not openly available because they are the subject of continuing research but are available from the first author on an appropriate request basis.

Declarations

Conflict of interest The authors have no conflicts of interest.

References

1. Allen JB (1977) Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Trans Acoust Speech Signal Process* ASSP-25:235–238
2. Bao G, Xu Y, Ye Z (2014) Learning a discriminative dictionary for single-channel speech separation. *IEEE/ACM Trans Audio Speech Lang Process* 22(7):1130–1138
3. Cooke M, Barker J, Cunningham S, Shao X (2006) An audio-visual corpus for speech perception and automatic speech recognition. *J Acoust Soc Am* 120(5):2421–2424
4. Demir C, Saraclar M, Cemgil A (2013) Single-channel speech-music separation for robust ASR with mixture models. *IEEE Trans Audio Speech Lang Process* 21(4):725–736
5. Fu J, Zhang L, Ye Z (2018) Supervised monaural speech enhancement using two level complementary joint sparse representations. *Appl Acoust* 132:1–7
6. Garofolo J et al (1993) TIMIT Acoustic-Phonetic Continuous Speech Corpus. LDC93S1, Web download, Philadelphia: Linguistic Data Consortium. <https://doi.org/10.35111/17gk-bn40>
7. Grais EM, Erdogan H (2013) Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation. In: *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)*, Lyon, France, pp. 808–812
8. Hossain MI, Islam MS, Khatun MT et al (2021) Dual-transform source separation using sparse nonnegative matrix factorization. *Circ Syst Signal Process* 40:1868–1891. <https://doi.org/10.1007/s00034-020-01564-x>
9. Huang PS, Kim M, Johnson MH, Smaragdis P (2015) Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Trans Audio Speech Lang Process* 23(12):2136–2147
10. Hurley N, Rickard S (2009) Comparing measures of sparsity. *IEEE Trans Inf Theory* 55(10):4723–4741

11. Islam MS, Al Mahmud TH, Khan WU, Ye Z (2019) Supervised single channel speech enhancement based on stationary wavelet transforms and nonnegative matrix factorization with concatenated framing process and subband smooth ratio mask. *J Sign Process Syst* 92:445–458. <https://doi.org/10.1007/s11265-019-01480-7>
12. Islam MS, Al Mahmud TH, Khan WU, Ye Z (2019) Supervised Single Channel speech enhancement based on dual-tree complex wavelet transforms and nonnegative matrix factorization using the joint learning process and subband smooth ratio mask. *Electronics* 8(3):353
13. Islam MS, Zhu YY, Hossain MI, Ullah R, Ye Z (2020) Supervised single channel dual domains speech enhancement using sparse non-negative matrix factorization. *Digital Signal Process* 100:102697
14. Islam MS, Naqvi N, Abbasi AT, Hossain MI, Ullah R, Khan R, Islam MS, Ye Z (2021) Robust dual domain twofold encrypted image-in-audio watermarking based on SVD. *Circ Syst Signal Process* 40:4651–4685
15. Jang GJ, Lee TW (2003) A maximum likelihood approach to single channel source separation. *J Mach Learn Res* 4:1365–1392
16. Jia H, Wang W, Wang Y, Pei J (2019) Speech enhancement based on discriminative joint sparse dictionary alternate optimization. *J Xidian Univ* 46(3):74–81
17. Jiang D, He Z, Lin Y, Chen Y, Xu L (2021) An improved unsupervised single-channel speech separation algorithm for processing speech sensor signals. *Wirel Commun Mob Comput* 2021. <https://doi.org/10.1155/2021/6655125>
18. Kates JM, Arehart KH (2010) The hearing-aid speech quality index (HASQI). *J Audio Eng Soc* 58(5):363–381
19. Kates JM, Arehart KH (2014) The hearing-aid speech perception index (HASPI). *Speech Comm* 65:75–93
20. Ke S, Hu R, Wang X, Wu T, Li G, Wang Z (2020) Single Channel multi-speaker speech separation based on quantized ratio mask and residual network. *Multimed Tools Appl* 79:32225–32241
21. Kingsbury NG (1998) The dual-tree complex wavelet transforms: a new efficient tool for image restoration and enhancement. In: *Proceedings of the 9th European Signal Process Conference, EUSIPCO, Rhodes, Greece*. pp. 319–322
22. Lee DD, Seung HS (1999) Learning the pans of objects with nonnegative matrix factorization. *Nature* 401:788–791
23. Lian Q, Shi G, Chen S (2015) Research progress of dictionary learning model, algorithm and its application. *J Autom* 41(2):240–260
24. Lorenz MO (1905) Methods of measuring concentrations of wealth. *J Am Stat Assoc* 9:209
25. Luo Y, Bao G, Xu Y, Ye Z (2015) Supervised monaural speech enhancement using complementary joint sparse representations. *IEEE Signal Process Lett* 23:237–241
26. Mowlae P, Saedi R, Christensen MG, Tan ZH, Kinnunen T, Franti P, Jensen SH (2012) A joint approach for single-channel speaker identification and speech separation. *IEEE Trans Audio Speech Lang Process* 20(9):2586–2601
27. Muhammed B, Lekshmi MS (2017) Single channel speech separation in transform domain combined with DWT. *National Conference on Technological Trends (NCTT)*, Manuscript Id: NCTTP006, pp. 15–18
28. Paatero P, Tapper U (1994) Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5(2):111–126
29. Rivet B, Wang W, Naqvi SM, Chambers JA (2014) Audiovisual speech source separation: an overview of key methodologies. *IEEE Signal Process Mag* 31(3):125–134
30. Rix A, Beerends J, Hollier M, Hekstra A (2010) Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. *IEEE International Conference on Acoustics, Speech, Signal Processing*, pp. 749–752
31. Roweis ST (2001) One microphone source separation. *Adv Neural Inf Process Syst* 13:793–799
32. Salman MS, Naqvi SM, Rehman A, Wang W, Chambers JA (2013) Video-aided model-based source separation in real reverberant rooms. *IEEE Trans Audio Speech Lang Process* 21(9):1900–1912
33. Sigg CD, Dikk T, Buhmann JM (2012) Speech enhancement using generative dictionary learning. *IEEE Trans Audio Speech Lang Process* 20(6):1698–1712
34. Sun Y, Rafique W, Chambers JA, Naqvi SM (2017) Undetermined source separation using time-frequency masks and an adaptive combined Gaussian-student's probabilistic model. In *Proc IEEE Int Conf Acoust Speech Signal Process* pp. 4187–4191
35. Sun L, Zhao C, Su M, Wang F (2018) Single-channel blind source separation based on joint dictionary with common sub-dictionary. *Int J Speech Technol* 21(1):19–27
36. Sun L, Xie K, Gu T, Chen J, Yang Z (2019) Joint dictionary learning using a new optimization method for single-channel blind source separation. *Speech Comm* 106:85–94
37. Sun Y, Xian Y, Wang W, Naqvi SM (2019) Monaural source separation in complex domain with long short-term memory neural network. *IEEE J Sel Top Signal Process* 13(2):359–369

38. Sun L, Zhu G, Li P (2020) Joint constraint algorithm based on deep neural network with dual outputs for single-channel speech separation. *SIViP* 14:1387–1395. <https://doi.org/10.1007/s11760-020-01676-6>
39. Sun L, Bu Y, Li P, Wu Z (2021) Single-channel speech enhancement based on joint constrained dictionary learning, Sun et al. *EURASIP J Audio Speech Music Process.* <https://doi.org/10.1186/s13636-021-00218-3>
40. Taal CH, Hendriks RC, Heusdens R, Jensen J (2011) An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans Audio Speech Lang Process* 19(7):2125–2136
41. Ullah R, Islam MS, Hossain MI, Wahab FE, Ye Z (2020) Single channel speech dereverberation and separation using RPCA and SNMF. *Appl Acoust* 167:107406. <https://doi.org/10.1016/j.apacoust.2020.107406>
42. Varshney YV, Abbasi ZA, Abidi MR, Farooq O (2017) Frequency selection based separation of speech signals with reduced computational time using sparse NMF. *Arch Acoust* 42(2):287–295
43. Vincent E, Gribonval R, Fevotte C (2006) Performance measurement in blind audio source separation. *IEEE Trans Audio Speech Lang Process* 14:1462–1469
44. Wang Y, Li Y, Ho KC, Zare A, Skubic M (2014) Sparsity promoted non-negative matrix factorization for source separation and detection. *Proceedings of the 19th International Conference on Digital Signal Processing.* IEEE, pp. 20–23
45. Wang Z, Sha F (2014) Discriminative nonnegative matrix factorization for Single-Channel speech separation. *IEEE International Conference on Acoustic, Speech and Signal Processing*
46. Williamson DS, Wang Y, Wang D (2016) Complex ratio masking for monaural speech separation. *IEEE/ACM Trans Audio Speech Lang Process* 24(3):483–492
47. Wu B, Li K, Yang M, Lee C-H (2017) A reverberation time aware approach to speech dereverberation based on deep neural networks. *IEEE/ACM Trans Audio Speech Lang Process* 25(1):102–111
48. Xu Y, Bao G, Xu X, Ye Z (2015) Single-channel speech separation using sequential discriminative dictionary learning. *Signal Process* 106:134–140
49. Yang M, Zhang L, Yang J, Zhang D (2010) Metaface learning for sparse representation based face recognition. *IEEE International Conference on Image Processing*, pp. 1601–1604
50. Zohrevandi M, Setayeshi S, Rabiee A et al (2021) Blind separation of underdetermined convolutive speech mixtures by time–frequency masking with the reduction of musical noise of separated signals. *Multimed Tools Appl* 80:12601–12618. <https://doi.org/10.1007/s11042-020-10398-3>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Md Imran Hossain received B.Sc. and M.Sc. degrees from the Information and Communication Engineering department of the Islamic University, Kushtia, Bangladesh in 2009 and 2010, respectively. He has been working as an Assistant Professor in the Department of Information and Communication Engineering of Pabna University of Science and Technology, Pabna, Bangladesh. Currently, he is pursuing a Ph.D. degree from the University of Science and Technology of China, Hefei, China. His current research interests are speech signal processing, such as speech separation, speech enhancement, and denoising.



Tarek Hasan Al Mahmud received B.Sc. and M.Sc. degrees in Information and Communication engineering from the Islamic University, Kushtia, Bangladesh, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2019. Currently, he is an associate professor department of ICE at the Islamic University, Kushtia, Bangladesh. His current research interests are array signal processing for high-resolution DOA estimation, and Speech Processing.



Md Shohidul Islam received a B.Sc. and M.Sc. degrees in Computer Science and Engineering from the Islamic University, Kushtia, Bangladesh in 2004 and 2005, respectively, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2020. Currently, he is an associate professor at the Islamic University, Kushtia, Bangladesh. His current research interests are Speech Signal processing, such as Speech Enhancement, Denoising, Speech Separation, and Dereverberation.



Md Bipul Hossen received the B.Sc. degree and the M.Sc. degree in Statistics department from Rajshahi University, Rajshahi, Bangladesh in 2011 and 2012 respectively. He is currently an Assistant professor in Statistics department at Begum Rokeya University, Rangpur, Bangladesh and doing his PhD degree under the Department of Electronic Engineering and Information Science in University of Science and Technology of China, Hefei, China. His research interests include statistical signal processing, machine learning, deep learning, and Bioinformatics.



Rashid Khan received the MS. Degree in Computer Engineering from Near East University Cyprus in 2014. He is currently doing his Ph.D. Degree from Dept. of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei China. His research activities are focused on Image Processing, Computer Vision, Natural language Processing, and Deep Learning.



Zhongfu Ye received the BE and MS degrees from the Hefei University of Technology, Hefei, China, in 1982 and 1986, respectively, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 1995. He is currently a Professor of the University of Science and Technology of China. His current research interests are in statistical and array signal processing and image processing.

Affiliations

Md Imran Hossain¹ · Tarek Hasan Al Mahmud² · Md Shohidul Islam³ · Md Bipul Hossen¹ · Rashid Khan¹ · Zhongfu Ye¹

Md Imran Hossain
imranpost@mail.ustc.edu.cn

Tarek Hasan Al Mahmud
tarek@mail.ustc.edu.cn

Md Shohidul Islam
shohid7@mail.ustc.edu.cn

Md Bipul Hossen
mbipu@mail.ustc.edu.cn

Rashid Khan
rashidkhan@mail.ustc.edu.cn

¹ National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei 230026 Anhui, China

² Department of ICE, Islamic University, Kushtia, Bangladesh

³ Department of CSE, Islamic University, Kushtia, Bangladesh