




Dual-Transform Source Separation Using Sparse Nonnegative Matrix Factorization

Md. Imran Hossain¹ · Md. Shohidul Islam² · Mst. Titasa Khatun³ · Rizwan Ullah¹ · Asim Masood¹ · Zhongfu Ye¹ 

Received: 2 December 2019 / Revised: 29 September 2020 / Accepted: 6 October 2020 /
Published online: 23 October 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

In this article, we propose a new source separation method in which the dual-tree complex wavelet transform (DTCWT) and short-time Fourier transform (STFT) algorithms are used sequentially as dual transforms and sparse nonnegative matrix factorization (SNMF) is used to factorize the magnitude spectrum. STFT-based source separation faces issues related to time and frequency resolution because it cannot exactly determine which frequencies exist at what time. Discrete wavelet transform (DWT)-based source separation faces a time-variation-related problem (i.e., a small shift in the time-domain signal causes significant variation in the energy of the wavelet coefficients). To address these issues, we utilize the DTCWT, which comprises two-level trees with different sets of filters and provides additional information for analysis and approximate shift invariance; these properties enable the perfect reconstruction of the time-domain signal. Thus, the time-domain signal is transformed into a set of subband signals in which low- and high-frequency components are isolated. Next, each subband is passed

✉ Zhongfu Ye
yefz@ustc.edu.cn

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

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

Mst. Titasa Khatun
titasa_ice09@yahoo.com

Rizwan Ullah
rizwanul@mail.ustc.edu.cn

Asim Masood
asimmasood@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 CSE, Islamic University, Kushtia 7003, Bangladesh
- ³ Department of ICE, Islamic University, Kushtia 7003, Bangladesh

through the STFT and a complex spectrogram is constructed. Then, SNMF is applied to decompose the magnitude part into a weighted linear combination of the trained basis vectors for both sources. Finally, the estimated signals can be obtained through a subband binary ratio mask by applying the inverse STFT (ISTFT) and the inverse DTCWT (IDTCWT). The proposed method is examined on speech separation tasks utilizing the GRID audiovisual and TIMIT corpora. The experimental findings indicate that the proposed approach outperforms the existing methods.

Keywords Speech separation (SS) · Dual-tree complex wavelet transform (DTCWT) · Sparse nonnegative matrix factorization (SNMF) · Short-time Fourier transform (STFT)

1 Introduction

Source separation (SS) is a procedure for isolating a set of source signals from an observed or mixed signal. Single-channel SS (SCSS) has become important in many real-world applications, such as communication, multimedia, and the cocktail-party problem. Although devices for SCSS have many obvious possible applications in hearing aids or as preprocessors in speech recognition systems, existing devices still show considerable room for improvement in performance. Research on SCSS for speech signals began a few decades ago [14, 20, 36] and continues to be conducted [21]. It is also a broadly examined issue in the machine learning community. Many signal models that consider numerous parameters (e.g., phase, magnitude, amplitude, frequency, energy, and the spectrogram of the speech signal) have been proposed.

S.T. Roweis suggested factorial hidden Markov models (HMMs), which have been incredibly successful for a single speaker [28, 29]. Jang and Lee [13] used a maximum likelihood approach to separate the mixed source signal that is perceived in a single channel. Pearlmutter and Olsson [24] exploited linear program differentiation on overcomplete dictionaries for maximally sparse representations of a speech corpus. Over the last several years, nonnegative matrix factorization (NMF) has become very popular with researchers for the separation of single-channel source signals. NMF was first introduced by Paatero and Tapper [3] and was proposed for use in SS by Lee and Seung [23]. NMF refers to a group of methods for multivariate analysis in which a matrix is decomposed into two other nonnegative matrices according to its components and weights.

Sliding windows and various types of mask-related NMF methods [4] have been used to decompose mixed signal magnitude spectra into weighted combinations of basis vectors for both sources. NMF-based algorithms are used to iteratively optimize a cost function [5]. Discriminative learning in NMF [41] is used to optimize all basis vectors jointly to reconstruct both clean and mixed signals. The magnitude spectrogram of a speech signal is primarily a two-dimensional matrix, and speech signals are typically sparse; thus, sparse NMF (SNMF) is applied to factorize them [9]. SCSS using SNMF was proposed by Schmidt and Olsson [30]. SNMF can learn a sparse representation of data [30] to solve the problem of separating multiple speech sources from a single microphone recording. SNMF enforces sparsity in both the basis matrix

and the coefficient matrix. For signal detection, however, sparsity is enforced in only the coefficient matrix [40]. Wang et al. [40] showed that the performance is improved with a suitable choice of basis vectors. Group sparse NMF with divergence and graph regularization [25] has been utilized to separate source signals. Variable sparsity regularization factor-based SNMF [43] has also been used to separate monaural speech signals.

Dictionary learning (DL)-based algorithms [1, 7, 33, 34, 42] are another effective class of methods for model-based SCSS. Sequential discriminative DL (SDDL) was presented in [42], where both the distinctive and similar parts of varying speaker signals were considered. The authors of [33] constructed a joint dictionary method with a common subdictionary (CJD) in which a common subdictionary was built using similar atoms between identity subdictionaries trained using source speech signals corresponding to each speaker, and these similar were then discarded from the identity subdictionaries. In [34], the authors proposed a new optimization function for preparing a joint dictionary with multiple identity subdictionaries and a common subdictionary.

Recently, the wavelet transform (WT) algorithm has been utilized in many different fields, for example, speech recognition [8], noise reduction [22, 26], and electrocardiography [32]. The authors of [37] proposed an improved model for separating the mixed speech signals. In this model, the high-frequency components of the signal are rejected to reduce the computation time, and the low-frequency components are separated by using the WT. Specifically, the signal is decomposed by using the discrete wavelet transform (DWT), and the signal coming from the highest 50% of the frequency band is considered noise and is replaced with zero. The DWT does not yield a good estimate of the critical subband decomposition because the high-frequency portion of the signal is fully rejected; hence, the performance of the speech separation process is degraded. The authors of [39] offered a speech enhancement (SE) approach that utilizes the discrete wavelet packet transform (DWPT) and provides adequate information both for analysis and synthesis of the original signal, with a remarkable reduction in computation time. The authors of [11] presented an SE method based on the stationary wavelet transform (SWT) and NMF that overcomes the time-variation-related problem. The authors of [12] offered another SE method with limited redundancy and time-invariant properties, which outperforms conventional methods at low signal-to-noise ratios (SNRs).

Short-time Fourier transform (STFT)-based SS faces problems in time and frequency resolution because it cannot precisely determine which frequencies exist at what time. DWT-based SS [37] faces time-variation-related issues that hamper its separation performance. The DWPT suffers from the shift-variance problem; i.e., small shifts in the input signal can cause large variations in the distribution of energy among coefficients at different levels, causing signal reproduction errors [39]. The SWT introduces redundancy-related problems [11]. Our proposed method can solve all of the above-mentioned issues to a certain extent.

In our work, we propose a dual-transform SS strategy in which the DTCWT is applied to decompose the time-domain signal and produce a set of subband signals. Then, the STFT is applied to each subband signal, to convert each subband signal into the time–frequency domain, and a complex spectrogram is built for each subband

signal. Finally, SNMF is applied to the magnitude spectrogram to obtain the weighted basis vectors to be used in the testing phase to separate the speech signals.

The contributions of this paper are briefly listed below:

- i. We first use the DTCWT to divide the input signal into small parts in order to separate the low- and high-frequency components. Then, the STFT is applied to each subband signal, which tends to be stationary and to provide a better transformation than other transforms. The sequential use of the DTCWT and STFT improves the separation capability of the model due to the approximate shift invariance and perfect reconstruction capabilities of the DTCWT.
- ii. SNMF is applied separately to the magnitude spectrogram of each subband signal to produce the weighted basis vectors that are then used during the testing process. The feature vectors can be more effectively extracted by applying SNMF to each subband signal.
- iii. Several WT- and STFT-based separation methods are compared, and our method is found to outperform the previous strategies mentioned in this paper.

The rest of the paper is organized as follows. Section 2 presents a mathematical description of the single-channel speech separation problem. Section 3 provides a brief explanation of the WT and SNMF methods. Section 4 presents the existing SS algorithms. Section 5 presents the details of the proposed algorithm. Section 6 describes the experimental setup and speech database and compares the experimental results. Section 7 concludes the presented work and is followed by the references. The nomenclature is provided in Table 1.

2 Problem Formulation

We consider two sources in our SS process, where the first source signal is $\mathbf{x}(t) = [x(1); x(2); \dots; x(T)]$, and the second source signal is $\mathbf{y}(t) = [y(1); y(2); \dots; y(T)]$; here, T and t denote the number of samples and the time instance, respectively. The mixed signal $\mathbf{z}(t)$ is prepared by summing the two source signals. The expression for the mixed signal is defined in Eq. (1).

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

Now, the DTCWT is applied to Eq. (1), as is shown in Eq. (2), to obtain the DTCWT subbands as presented in Eq. (3).

$$DTCWT\{\mathbf{z}(t)\} = DTCWT\{\mathbf{x}(t)\} + DTCWT\{\mathbf{y}(t)\} \quad (2)$$

$$\mathbf{z}_{b,tl}^J = \mathbf{x}_{b,tl}^J + \mathbf{y}_{b,tl}^J \quad (3)$$

where $\mathbf{z}_{b,tl}^J$, $\mathbf{x}_{b,tl}^J$, and $\mathbf{y}_{b,tl}^J$ represent the mixed, first source, and second source subband signals, respectively, and J , b , and tl denote the level of the DTCWT, the subband index,

Table 1 Nomenclature

Symbols	Abbreviations
x, X (lowercase and uppercase)	Variables
\mathbf{x} (lowercase bold)	Vector
\mathbf{X} (uppercase bold)	Matrix
X (uppercase italic)	Function
\mathbf{X} (uppercase bold italic)	Method
\otimes	Elementwise multiplication
$\sqrt{\cdot}$	Elementwise square root operation
SE	Speech enhancement
SS	Source separation
STFT	Short-time Fourier transform
ISTFT	Inverse short-time Fourier transform
NMF	Nonnegative matrix factorization
SNMF	Sparse nonnegative matrix factorization
DWT	Discrete wavelet transform
IDWT	Inverse discrete wavelet transform
DWPT	Discrete wavelet packet transform
IDWPT	Inverse discrete wavelet packet transform
SWT	Stationary wavelet transform
DTCWT	Dual-tree complex wavelet transform
IDTCWT	Inverse dual-tree complex wavelet transform
CJD	Joint dictionary method with a common subdictionary
JDL	Joint dictionary learning
SNR	Signal-to-noise ratio
KL	Kullback–Leibler
FB	Filter bank
PM	Proposed method
SBRMX	Subband binary ratio mask of signal x
SBRMY	Subband binary ratio mask of signal y
STFT – SNMF	STFT- and SNMF-based SS method [41]
DTCWT – SNMF	DTCWT- and SNMF-based SS method followed by DTCWT – NMF [12]
DWT – STFT – SNMF	DWT-, STFT-, and SNMF-based SS method [37]
DWPT – SNMF	DWPT- and SNMF-based SS method followed by DWPT – NMF [39]
SWT – SNMF	SWT- and SNMF-based SS method followed by SWT – NMF [22]
DTCWT – STFT – SNMF	DTCWT-, STFT-, and SNMF-based SS method [PM]
HASQI	Hearing-Aid Speech Quality Index [16]

Table 1 continued

Symbols	Abbreviations
HASPI	Hearing-Aid Speech Perception Index [15]
PESQ	Perceptual evaluation of speech quality [27]
STOI	Short-time objective intelligibility [35]
fwsegSNR	Average frequency-weighted segmental SNR [38]
SDR	Source distortion ratio [10]
SIR	Signal-to-interference ratio [10]

and the tree level, respectively. The STFT of Eq. (3) is defined in Eq. (4) and yields the complex matrices represented in Eq. (5).

$$STFT\{z_{b,t}^J\} = STFT\{x_{b,t}^J\} + STFT\{y_{b,t}^J\} \quad (4)$$

$$\mathbf{Z}_{b,t}^J(\tau, f) = \mathbf{X}_{b,t}^J(\tau, f) + \mathbf{Y}_{b,t}^J(\tau, f) \quad (5)$$

where $\mathbf{Z}_{b,t}^J(\tau, f)$, $\mathbf{X}_{b,t}^J(\tau, f)$, and $\mathbf{Y}_{b,t}^J(\tau, f)$ are the STFT coefficients of $\mathbf{z}_{b,t}^J$, $\mathbf{x}_{b,t}^J$, and $\mathbf{y}_{b,t}^J$, respectively, and f and τ are the frequency and time bin indexes, respectively. $\tilde{\mathbf{X}}_{b,t}^J(\tau, f)$ and $\tilde{\mathbf{Y}}_{b,t}^J(\tau, f)$ are the unknown complex coefficients for the sources and must be estimated via SNMF, the subband binary ratio mask of signal \mathbf{x} (SBRMX), and the subband binary ratio mask of signal \mathbf{y} (SBRMY) from Eq. (5) using only the magnitude part. Finally, the estimated first and second source signals are calculated via the following equations:

$$\tilde{\mathbf{x}}_{b,t}^J = ISTFT(\tilde{\mathbf{X}}_{b,t}^J(\tau, f)) \quad (6)$$

$$\tilde{\mathbf{y}}_{b,t}^J = ISTFT(\tilde{\mathbf{Y}}_{b,t}^J(\tau, f)) \quad (7)$$

$$\tilde{\mathbf{x}}(t) = IDTCWT(\tilde{\mathbf{x}}_{b,t}^J) \quad (8)$$

$$\tilde{\mathbf{y}}(t) = IDTCWT(\tilde{\mathbf{y}}_{b,t}^J) \quad (9)$$

where $\tilde{\mathbf{x}}(t)$ and $\tilde{\mathbf{y}}(t)$ are the estimated first and second source signals, respectively, and ISTFT and IDTCWT denote the inverse STFT and inverse DTCWT, respectively.

3 Review of the WT and SNMF

The WT algorithm generates an assortment of time–frequency representations of a signal with various resolutions. The WT utilizes a completely versatile adjusted window that is moved along the signal, and the spectrum is computed for every position.

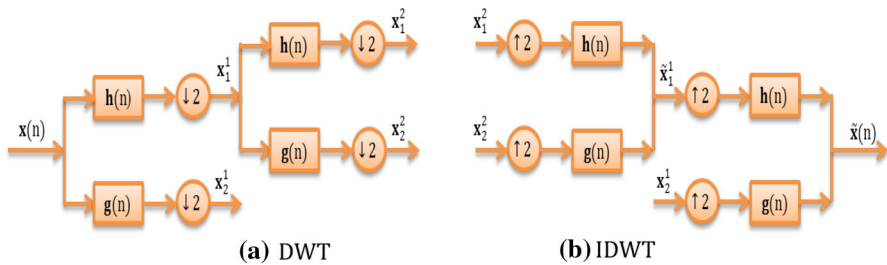


Fig. 1 Second-level block diagrams of (a) the DWT and (b) the IDWT

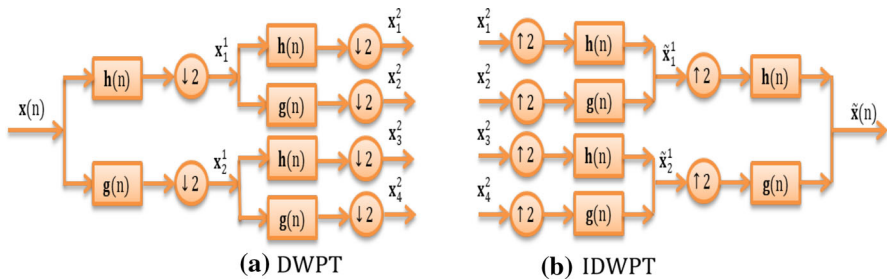


Fig. 2 Second-level block diagrams of (a) the DWPT and (b) the IDWPT

Recently, the DWT has been effectively adopted in many signal processing applications because of its ability to provide an efficient time–frequency analysis of a signal. The DWT decomposes the time-domain signal $x(n)$ using a pair of low-pass ($h(n)$) and high-pass ($g(n)$) filters, and the signal is then downsampled by a factor of two. The output of the low-pass filter is known as an approximation coefficient (x_1^1), where the superscript and subscript of x denote the DWT level and subband index, respectively, and represent the high-frequency part of the signal. The output of the high-pass filter is called the detail coefficient (x_2^1) and represents the low-frequency part of the signal. For the next level of decomposition, only x_1^1 is passed through similar low-pass and high-pass filters and is then downsampled to obtain x_1^2 and x_2^2 , and so on. Figure 1 illustrates the filter bank (FB) implementations of the DWT and IDWT.

The DWPT is a generalized variant of the DWT. At the first level, the DWPT decomposes the time-domain signal $x(n)$ using a pair of low-pass ($h(n)$) and high-pass ($g(n)$) filters and performs downsampling by a factor of two. In second-level decomposition, both the approximation and detail coefficients are subjected to similar low-pass and high-pass filters and downsampling to obtain the approximation and detail coefficients at the next level, and so on. Figure 2 illustrates the FB implementations of the DWPT and IDWPT. Although the DWT and DWPT have practical computational advantages, they also have some drawbacks, such as shift variance, oscillation, aliasing, and a lack of directionality.

In the SWT, the downsampling process after filtration at each level is removed, and consequently, the coefficient length is the same as the length of the original time-domain signal. In the SWT, the time-domain signal is passed through the high-pass and low-pass filters at the first level to obtain the corresponding approximation and

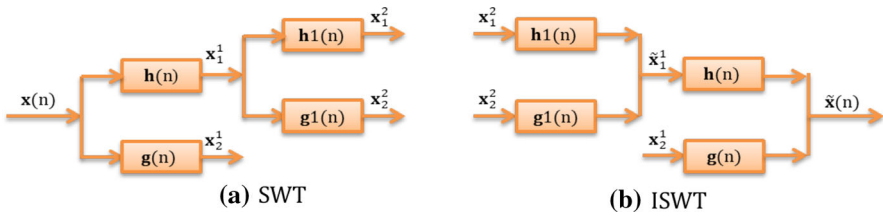


Fig. 3 Second-level block diagrams of (a) the SWT and (b) the ISWT

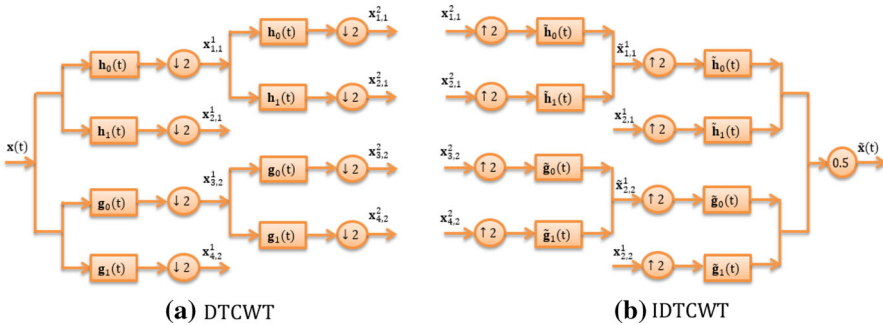


Fig. 4 Second-level block diagrams of (a) the DTCWT and (b) the IDTCWT

detail coefficients. Then, at the second level, the low-pass and high-pass filters are upsampled by adding a zero between each pair of the adjacent filter of components from the previous level, and only the approximation coefficient is passed through the newly updated low-pass and high-pass filters; a similar process is repeated at each subsequent level. Thus, the SWT eliminates shift-invariance issues by disposing of downsampling operators; however, it introduces redundancy. Figure 3 shows the second-level block diagrams of the SWT and ISWT.

To mitigate the issues of shift variance and redundancy simultaneously, Kingsbury [17] presented a computationally advantageous method called the DTCWT, which is shift-invariant and has limited redundancy. In the first level, it decomposes the time-domain signal into four subband signals corresponding to two trees, where the first tree provides the real part of the transform, while the second tree provides the imaginary part. The upper and lower trees each yield one approximation coefficient ($x_{1,1}^1$) and one detail coefficient ($x_{2,1}^1$), where the superscript denotes the DTCWT level, and the first and second subscripts represent the subband index and tree level, respectively. Then, each subband signal is downsampled. In the second level of decomposition, only the approximation coefficients are passed through the filters to produce second-level subband signals, and so on. Figure 4 illustrates the FB implementation of the DTCWT and IDTCWT.

NMF is an algorithm for analysis in which a matrix of nonnegative elements is factorized into two nonnegative matrices according to its bases and weights. In the factorization process, the matrix $Z \in \mathbb{R}^{F \times T}$ is decomposed as a linear combination

of bases $\mathbf{W} \in \mathbb{R}^{F \times R}$ and weights $\mathbf{H} \in \mathbb{R}^{R \times T}$, where the inner dimension R is much smaller than the product of F and T of the original matrix \mathbf{Z} :

$$\mathbf{Z} \approx \mathbf{W}\mathbf{H} \quad (10)$$

NMF has been a popular method for modeling speech signals, especially in single-channel speech separation applications. Regardless of the size of the corpus, it will not be possible to learn a substantial dictionary with more elements than the number of time–frequency bins [18]. To address this issue, the authors of [31] presented sparsity penalties on the activations \mathbf{H} concerning audio. Additionally, cost functions based on the Euclidean distance and Kullback–Leibler (KL) divergence were investigated in [19], and it was shown that the KL cost function works outstandingly well for audio SS. Therefore, we consider SNMF with the KL cost function in our study. The KL divergence cost function is minimized as follows:

$$\begin{aligned} C_{\text{KL}} &= \min D(\mathbf{Z}||\mathbf{W}\mathbf{H}) + \mu \|\mathbf{H}\|_1 \\ &= \sum_{i,j} (\mathbf{Z}_{i,j} \log \frac{\mathbf{Z}_{i,j}}{(\mathbf{W}\mathbf{H})_{i,j}} - \mathbf{Z}_{i,j} + (\mathbf{W}\mathbf{H})_{i,j}) + \mu \sum_{i,j} |\mathbf{H}_{i,j}| \end{aligned} \quad (11)$$

where μ denotes the sparsity parameter. The matrices \mathbf{W} and \mathbf{H} are updated through the following iterative principles:

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\frac{\mathbf{Z}}{\mathbf{W}\mathbf{H}}\mathbf{H}^T + \mathbf{W} \otimes (1_v(\sum(\mathbf{W} \otimes 1_m\mathbf{H}^T)))}{1_m\mathbf{H}^T + \mathbf{W} \otimes (1_v(\sum(\mathbf{W} \otimes (\frac{\mathbf{Z}}{\mathbf{W}\mathbf{H}}\mathbf{H}^T))))} \quad (12)$$

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T \frac{\mathbf{Z}}{\mathbf{W}\mathbf{H}}}{\mathbf{W}^T 1_m + \mu} \quad (13)$$

where 1_m denotes a matrix of ones with the same dimensions as of \mathbf{Z} , 1_v denotes a column vector of ones with a number of entries equal to the number of columns of \mathbf{W} , and all of the divisions' operations elementwise division.

4 SS and SE Algorithms Based on NMF or SNMF

The STFT – SNMF-based SS algorithm [41] has a training stage and testing stage. In the training stage, the STFT is applied to the clean speech signals $\mathbf{x}(t)$ and $\mathbf{y}(t)$ to produce complex-valued spectrograms $\mathbf{X}(\tau, f)$ and $\mathbf{Y}(\tau, f)$, respectively, where τ is the time index and f is the frequency index. From these complex spectrograms, only the magnitude spectra $|\mathbf{X}(\tau, f)|$ and $|\mathbf{Y}(\tau, f)|$ are considered and are passed through the SNMF module to obtain clean speech spectral bases \mathbf{W}_X and \mathbf{W}_Y . Finally, these spectral bases are concatenated to form $\mathbf{W}_{XY} = [\mathbf{W}_X \mathbf{W}_Y]$, which is used to prepare the activation matrix \mathbf{H}_{XY} for the testing stage. In the testing stage, the STFT is applied to the mixed speech signal $\mathbf{z}(t)$, to produce a complex-valued mixed speech spectrogram $\mathbf{Z}(\tau, f)$. Then, only the magnitude spectrum $|\mathbf{Z}(\tau, f)|$ and the previously

formed spectral basis \mathbf{W}_{XY} are passed to the SNMF module to update the activation matrix $\mathbf{H}_{XY} = [\mathbf{H}_X \mathbf{H}_Y]$. Finally, the primary estimated clean speech signal spectra $|\tilde{\mathbf{X}}(\tau, f)|$ and $|\tilde{\mathbf{Y}}(\tau, f)|$ are obtained using Eqs. (14) and (15).

$$|\tilde{\mathbf{X}}(\tau, f)| = \mathbf{W}_X \mathbf{H}_X \tag{14}$$

$$|\tilde{\mathbf{Y}}(\tau, f)| = \mathbf{W}_Y \mathbf{H}_Y \tag{15}$$

The DWT–STFT–SNMF-based SS algorithm was presented in [37]. In the training stage, clean speech bases (dictionaries) are estimated from clean speech databases by applying SNMF after the DWT and STFT. In the testing stage, the DWT is applied to decompose the mixed speech signal $\mathbf{z}(t)$ into a set of coefficients consisting of one approximation coefficient and J detail coefficients (where J is the final decomposition level index). The STFT is used to decompose the approximation coefficient of the mixed speech signal to produce a complex-valued mixed speech spectrogram $\mathbf{Z}(\tau, f)$, where τ and f are the time index and frequency bin indexes, respectively, and all of the detail coefficients are replaced with zeros of the same length. Then, SNMF is used to factorize only the magnitude part of the complex mixed speech spectrogram, while the phase spectrogram is preserved for further processing. The clean speech spectrograms $\mathbf{X}(\tau, f)$ and $\mathbf{Y}(\tau, f)$ are estimated by using the corresponding bases and weights. Finally, the estimated clean time-domain speech signals $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are obtained by using the ISTFT and IDWT.

The DWPT – NMF-based SE algorithm was presented in [39]. In this algorithm, the training stage is the same as in the previous method except that the DWPT is used instead of the DWT (i.e., the bases \mathbf{W}_X^b and \mathbf{W}_N^b are estimated from the clean speech signal $\mathbf{x}(t)$ and the noise $\mathbf{n}(t)$, and the basis matrix $\mathbf{W}_{XN}^b = [\mathbf{W}_X^b \mathbf{W}_N^b]$ is concatenated). In the second phase, the DWPT and overlapping framing patterns are applied to the noisy speech signal $\mathbf{z}(t)$, and a set of nonnegative matrices \mathbf{Z}_b^J is obtained for the noisy speech signal. Then, the weights \mathbf{H}_X^b and \mathbf{H}_N^b for the speech and noise signals, respectively, are obtained by performing NMF on the matrix \mathbf{Z}_b^J using the fixed training basis matrix $\mathbf{W}_{XN}^b = [\mathbf{W}_X^b \mathbf{W}_N^b]$ and the initial random weight matrix $\mathbf{H}_{XN}^b = \begin{bmatrix} \mathbf{H}_X^b \\ \mathbf{H}_N^b \end{bmatrix}$. The gain sequence \mathbf{G}_b^J and the estimated subband signal $\tilde{\mathbf{x}}_b^J$ are obtained by using Eqs. (16) and (17), as follows:

$$\mathbf{G}_b^J = \sqrt{(\mathbf{W}_X^b \mathbf{H}_X^b \cdot / (\mathbf{W}_X^b \mathbf{H}_X^b + \mathbf{W}_N^b \mathbf{H}_N^b))} \tag{16}$$

$$\tilde{\mathbf{x}}_b^J = \mathbf{x}_b^J \otimes \mathbf{g}_b^J \tag{17}$$

The estimated subband signal $\tilde{\mathbf{x}}_b^J$ is generated by utilizing Eq. (18), where $\sigma_{b,c}$ and σ_b denote the root-mean-square values of the clean speech signal and $\tilde{\mathbf{x}}_b^J$, respectively. Finally, the de-framing scheme and the IDWPT are applied to obtain the estimated signal $\tilde{\mathbf{x}}(t)$.

$$\tilde{\mathbf{x}}_b^J = \frac{\sigma_{b,c}}{\sigma_b} \mathbf{x}_b^J \tag{18}$$

The SWT – NMF SE method was presented in [11]. In the training stage, the nonnegative matrices $\mathbf{X}_b^{J_{\text{Train}}}$ and $\mathbf{N}_b^{J_{\text{Train}}}$ are obtained from the clean speech signal $\mathbf{x}(t)$ and the noise $\mathbf{n}(t)$ using the SWT, the overlapping framing scheme, the concatenated framing process, and an autoregressive moving average filter, where J denotes the SWT level, and b denotes the subband index. Then, the basis matrices \mathbf{W}_X^b and \mathbf{W}_N^b that are obtained after NMF are concatenated to prepare the basis matrix $\mathbf{W}_{XN}^b = [\mathbf{W}_X^b \mathbf{W}_N^b]$. In the testing stage, rough estimates of the clean speech signal ($\bar{\mathbf{X}}_b^J$) and the noise ($\bar{\mathbf{N}}_b^J$) are calculated using Eqs. (19) and (20) after applying the SWT to the noisy speech signal $\mathbf{z}(t)$ and performing NMF on $\mathbf{Z}_b^{J_{\text{Test}}}$. Finally, the estimated time-domain signal $\bar{\mathbf{x}}(t)$ is obtained through the inverse concatenated framing process and the ISWT.

$$\bar{\mathbf{X}}_b^J = \mathbf{W}_X^b \mathbf{H}_X^b \quad (19)$$

$$\bar{\mathbf{N}}_b^J = \mathbf{W}_N^b \mathbf{H}_N^b \quad (20)$$

A recent study [12] proposed the DTCWT – NMF SE strategy, which utilizes the DTCWT, NMF with the KL cost function, and a proposed subband smooth ratio mask. We implement the DWPT – SNMF, SWT – SNMF, and DTCWT – SNMF methods for SS, analogously to the DWPT – NMF [11, 39] and DTCWT – NMF [12] methods for SE, respectively, and compare the results with those of the proposed method.

5 Proposed SS Algorithm

This section describes the newly proposed SS method and the subtleties related to this approach. Usually, speech signals have some high-frequency components and some low-frequency components. The low-frequency components of a signal contain a substantial amount of information, whereas the high-frequency components contain a negligible amount of information. Nevertheless, the high-frequency information impacts the basis vectors of the lower-frequency components. For this reason, SS using only NMF cannot properly estimate the contents of mixed speech signals. Therefore, the DTCWT is used to filter out the high- and low-frequency components of the signal. In our proposed method, we use the first-level decomposition, in which the time-domain signal is decomposed into four subband signals. For example, the DTCWT decomposes the source signal $\mathbf{x}(t)$ into components, denoted by $\mathbf{x}_{b,tl}^J$, where J denotes the DTCWT level, b is the subband index, and tl represents tree level. For the first-level decomposition, $J = 1$; $b = 1, 2, 3,$ and 4 ; and $tl = 1,$ and 2 , where 1 is for the upper tree and 2 is for the lower tree (i.e., the subbands are $\mathbf{x}_{1,1}^1, \mathbf{x}_{2,1}^1, \mathbf{x}_{3,2}^1,$ and $\mathbf{x}_{4,2}^1$, as explained in the DTCWT part of Sect. 3).

Then, the STFT is applied to each subband signal to convert each subband signal into the time–frequency domain and build a complex spectrogram for each subband signal. The STFT suffers from issues regarding time and frequency resolution because it cannot exactly determine which frequencies exist at what time. In our proposed method, we use the DTCWT and STFT sequentially to solve this problem of traditional STFT-based methods. First, we use the DTCWT to isolate the input signal

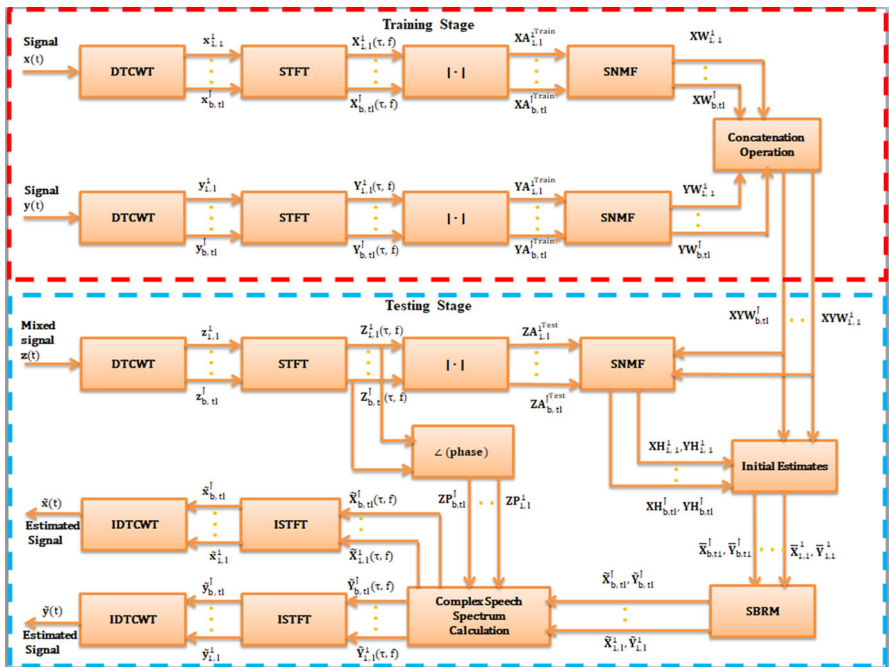


Fig. 5 Block diagram of the proposed SS approach

into sufficiently small portions such that the low- and high-frequency components are separated. The sequential use of the DTCWT and STFT makes the input signal more stationary, thus resulting in a better transformation. Finally, SNMF is applied to only the magnitude spectrum in order to factorize the bases and weight vectors. Figure 5 shows the overall block diagram of the SS algorithm proposed in this paper. The proposed algorithm is divided into two stages: a training stage and a testing stage.

5.1 Training Stage

The two signals $x(t)$ and $y(t)$ are utilized as source signals. We obtain the subband source signals $x_{1,1}^1, \dots, x_{b,tl}^J$ and $y_{1,1}^1, \dots, y_{b,tl}^J$ from the signals $x(t)$ and $y(t)$ by utilizing the DTCWT, where $J, b,$ and tl denote the DTCWT level, the subband index, and the tree level, respectively. The complex spectrograms $X_{1,1}^1(\tau, f), \dots, X_{b,tl}^J(\tau, f)$ and $Y_{1,1}^1(\tau, f), \dots, Y_{b,tl}^J(\tau, f)$ are calculated from the subband source signals via the STFT, where τ and f are the time and frequency bin indexes, respectively. The magnitude spectra $XA_{1,1}^{1Train}, \dots, XA_{b,tl}^{JTrain}$ and $YA_{1,1}^{1Train}, \dots, YA_{b,tl}^{JTrain}$ are obtained using Eqs. (21) and (22) through SNMF.

$$XA_{1,1}^{1Train}, \dots, XA_{b,tl}^{JTrain} \approx XW_{1,1}^1 \mathbf{X}H_{1,1}^1 + \mu \left| \mathbf{X}H_{1,1}^1 \right|, \dots, XW_{b,tl}^J \mathbf{X}H_{b,tl}^J + \mu \left| \mathbf{X}H_{b,tl}^J \right| \quad (21)$$

$$\mathbf{Y}\mathbf{A}_{1,1}^{\text{Train}}, \dots, \mathbf{Y}\mathbf{A}_{b,t}^{\text{Train}} \approx \mathbf{Y}\mathbf{W}_{1,1}^1 \mathbf{Y}\mathbf{H}_{1,1}^1 + \mu \left| \mathbf{Y}\mathbf{H}_{1,1}^1 \right|_1, \dots, \mathbf{Y}\mathbf{W}_{b,t}^J \mathbf{Y}\mathbf{H}_{b,t}^J + \mu \left| \mathbf{Y}\mathbf{H}_{b,t}^J \right|_1 \tag{22}$$

where $\mathbf{X}\mathbf{W}_{1,1}^1, \dots, \mathbf{X}\mathbf{W}_{b,t}^J$ and $\mathbf{X}\mathbf{H}_{1,1}^1, \dots, \mathbf{X}\mathbf{H}_{b,t}^J$ denote the basis and weight matrices for source signal one, $\mathbf{Y}\mathbf{W}_{1,1}^1, \dots, \mathbf{Y}\mathbf{W}_{b,t}^J$ and $\mathbf{Y}\mathbf{H}_{1,1}^1, \dots, \mathbf{Y}\mathbf{H}_{b,t}^J$ denote the basis and weight matrices for source signal two, and μ is the sparsity constant. The basis matrices $\mathbf{X}\mathbf{W}_{1,1}^1, \dots, \mathbf{X}\mathbf{W}_{b,t}^J$ can be generated by minimizing the distance between $\mathbf{X}\mathbf{A}_{1,1}^{\text{Train}}, \dots, \mathbf{X}\mathbf{A}_{b,t}^{\text{Train}}$ and $\mathbf{X}\mathbf{W}_{1,1}^1 \mathbf{X}\mathbf{H}_{1,1}^1 + \mu \left| \mathbf{X}\mathbf{H}_{1,1}^1 \right|_1, \dots, \mathbf{X}\mathbf{W}_{b,t}^J \mathbf{X}\mathbf{H}_{b,t}^J + \mu \left| \mathbf{X}\mathbf{H}_{b,t}^J \right|_1$ using Eq. (11), with the help of Eqs. (12) and (13). The basis matrices $\mathbf{Y}\mathbf{W}_{1,1}^1, \dots, \mathbf{Y}\mathbf{W}_{b,t}^J$ are generated similarly and are then concatenated with $\mathbf{X}\mathbf{W}_{1,1}^1, \dots, \mathbf{X}\mathbf{W}_{b,t}^J$ as follows: $\mathbf{X}\mathbf{Y}\mathbf{W}_{1,1}^1, \dots, \mathbf{X}\mathbf{Y}\mathbf{W}_{b,t}^J = [\mathbf{X}\mathbf{W}_{1,1}^1 \ \mathbf{Y}\mathbf{W}_{1,1}^1], \dots, [\mathbf{X}\mathbf{W}_{b,t}^J \ \mathbf{Y}\mathbf{W}_{b,t}^J]$.

5.2 Testing Stage

The mixed speech signal $z(t)$ is decomposed by applying the DTCWT to generate a set of subband signals $z_{1,1}^1, \dots, z_{b,t}^J$. The complex spectrum $\mathbf{Z}_{1,1}^1(\tau, f), \dots, \mathbf{Z}_{b,t}^J(\tau, f)$ is obtained from the individual subband signals using the STFT. The magnitude spectrum $\mathbf{Z}\mathbf{A}_{1,1}^{\text{Test}}, \dots, \mathbf{Z}\mathbf{A}_{b,t}^{\text{Test}}$ and phase spectrum $\mathbf{Z}\mathbf{P}_{1,1}^1, \dots, \mathbf{Z}\mathbf{P}_{b,t}^J$ are measured from the complex spectrum $\mathbf{Z}_{1,1}^1(\tau, f), \dots, \mathbf{Z}_{b,t}^J(\tau, f)$ by taking the absolute value and angle, respectively. The magnitude spectrum $\mathbf{Z}\mathbf{A}_{1,1}^{\text{Test}}, \dots, \mathbf{Z}\mathbf{A}_{b,t}^{\text{Test}}$ is decomposed via the SNMF using Eq. (23):

$$\begin{aligned} \mathbf{Z}\mathbf{A}_{1,1}^{\text{Test}}, \dots, \mathbf{Z}\mathbf{A}_{b,t}^{\text{Test}} &\approx \mathbf{X}\mathbf{Y}\mathbf{W}_{1,1}^1 \mathbf{X}\mathbf{Y}\mathbf{H}_{1,1}^1 + \mu \left| \mathbf{X}\mathbf{Y}\mathbf{H}_{1,1}^1 \right|_1, \dots, \mathbf{X}\mathbf{Y}\mathbf{W}_{b,t}^J \mathbf{X}\mathbf{Y}\mathbf{H}_{b,t}^J + \mu \left| \mathbf{X}\mathbf{Y}\mathbf{H}_{b,t}^J \right|_1 \\ &= [\mathbf{X}\mathbf{W}_{1,1}^1 \ \mathbf{Y}\mathbf{W}_{1,1}^1] \begin{bmatrix} \mathbf{X}\mathbf{H}_{1,1}^1 \\ \mathbf{Y}\mathbf{H}_{1,1}^1 \end{bmatrix} + \mu \left| \mathbf{X}\mathbf{H}_{1,1}^1 \ \mathbf{Y}\mathbf{H}_{1,1}^1 \right|_1, \dots, [\mathbf{X}\mathbf{W}_{b,t}^J \ \mathbf{Y}\mathbf{W}_{b,t}^J] \begin{bmatrix} \mathbf{X}\mathbf{H}_{b,t}^J \\ \mathbf{Y}\mathbf{H}_{b,t}^J \end{bmatrix} + \mu \left| \left[\mathbf{X}\mathbf{H}_{b,t}^J \ \mathbf{Y}\mathbf{H}_{b,t}^J \right] \right|_1 \end{aligned} \tag{23}$$

where $\mathbf{X}\mathbf{Y}\mathbf{H}_{1,1}^1, \dots, \mathbf{X}\mathbf{Y}\mathbf{H}_{b,t}^J, \mathbf{X}\mathbf{H}_{1,1}^1, \dots, \mathbf{X}\mathbf{H}_{b,t}^J$, and $\mathbf{Y}\mathbf{H}_{1,1}^1, \dots, \mathbf{Y}\mathbf{H}_{b,t}^J$ denote the weight matrices of the mixed signal, source signal one, and source signal two, respectively. The weight matrices $\mathbf{X}\mathbf{Y}\mathbf{H}_{1,1}^1, \dots, \mathbf{X}\mathbf{Y}\mathbf{H}_{b,t}^J$ can be learned via SNMF by minimizing the distances between $\mathbf{Z}\mathbf{A}_{1,1}^{\text{Test}}, \dots, \mathbf{Z}\mathbf{A}_{b,t}^{\text{Test}}$ and $\mathbf{X}\mathbf{Y}\mathbf{W}_{1,1}^1 \mathbf{X}\mathbf{Y}\mathbf{H}_{1,1}^1 + \mu \left| \mathbf{X}\mathbf{Y}\mathbf{H}_{1,1}^1 \right|_1, \dots, \mathbf{X}\mathbf{Y}\mathbf{W}_{b,t}^J \mathbf{X}\mathbf{Y}\mathbf{H}_{b,t}^J + \mu \left| \mathbf{X}\mathbf{Y}\mathbf{H}_{b,t}^J \right|_1$ using Eq. (11) with the help of Eq. (13), where the initial values of $\mathbf{X}\mathbf{Y}\mathbf{H}_{1,1}^1, \dots, \mathbf{X}\mathbf{Y}\mathbf{H}_{b,t}^J$ are initialized as random numbers, and the values of $\mathbf{X}\mathbf{Y}\mathbf{W}_{1,1}^1, \dots, \mathbf{X}\mathbf{Y}\mathbf{W}_{b,t}^J$ are fixed. The initially estimated magnitudes for source signal one ($\bar{\mathbf{X}}_{1,1}^1, \dots, \bar{\mathbf{X}}_{b,t}^J$) and source signal two ($\bar{\mathbf{Y}}_{1,1}^1, \dots, \bar{\mathbf{Y}}_{b,t}^J$) are obtained using Eqs. (24) and (25).

$$\bar{\mathbf{X}}_{1,1}^1, \dots, \bar{\mathbf{X}}_{b,t}^J = \mathbf{X}\mathbf{W}_{1,1}^1 \mathbf{X}\mathbf{H}_{1,1}^1, \dots, \mathbf{X}\mathbf{W}_{b,t}^J \mathbf{X}\mathbf{H}_{b,t}^J \tag{24}$$

$$\bar{\mathbf{Y}}_{1,1}^1, \dots, \bar{\mathbf{Y}}_{b,t}^J = \mathbf{Y}\mathbf{W}_{1,1}^1 \mathbf{Y}\mathbf{H}_{1,1}^1, \dots, \mathbf{Y}\mathbf{W}_{b,t}^J \mathbf{Y}\mathbf{H}_{b,t}^J \tag{25}$$

We observe that the sums of the initial estimates $\bar{\mathbf{X}}_{1,1}^1, \dots, \bar{\mathbf{X}}_{b,t}^J$ and $\bar{\mathbf{Y}}_{1,1}^1, \dots, \bar{\mathbf{Y}}_{b,t}^J$ are not equal to the components of the mixed signal magnitude spectrum $\mathbf{Z}\mathbf{A}_{1,1}^{1\text{Test}}, \dots, \mathbf{Z}\mathbf{A}_{b,t}^{J\text{Test}}$. To eliminate errors, we calculate the subband ratio masks using Eqs. (26) and (27).

$$\mathbf{SBRMX}_{1,1}^1, \dots, \mathbf{SBRMX}_{b,t}^J = \frac{(\bar{\mathbf{X}}_{1,1}^1)^2}{(\bar{\mathbf{X}}_{1,1}^1)^2 + (\bar{\mathbf{Y}}_{1,1}^1)^2}, \dots, \frac{(\bar{\mathbf{X}}_{b,t}^J)^2}{(\bar{\mathbf{X}}_{b,t}^J)^2 + (\bar{\mathbf{Y}}_{b,t}^J)^2} \tag{26}$$

$$\mathbf{SBRMY}_{1,1}^1, \dots, \mathbf{SBRMY}_{b,t}^J = \frac{(\bar{\mathbf{Y}}_{1,1}^1)^2}{(\bar{\mathbf{X}}_{1,1}^1)^2 + (\bar{\mathbf{Y}}_{1,1}^1)^2}, \dots, \frac{(\bar{\mathbf{Y}}_{b,t}^J)^2}{(\bar{\mathbf{X}}_{b,t}^J)^2 + (\bar{\mathbf{Y}}_{b,t}^J)^2} \tag{27}$$

The estimated source signal magnitudes $\tilde{\mathbf{X}}_{1,1}^1, \dots, \tilde{\mathbf{X}}_{b,t}^J$ and $\tilde{\mathbf{Y}}_{1,1}^1, \dots, \tilde{\mathbf{Y}}_{b,t}^J$ are obtained by using Eqs. (28) and (29).

$$\tilde{\mathbf{X}}_{1,1}^1, \dots, \tilde{\mathbf{X}}_{b,t}^J = \mathbf{Z}\mathbf{A}_{1,1}^{1\text{Test}} \otimes \mathbf{SBRMX}_{1,1}^1, \dots, \mathbf{Z}\mathbf{A}_{b,t}^{J\text{Test}} \otimes \mathbf{SBRMX}_{b,t}^J \tag{28}$$

$$\tilde{\mathbf{Y}}_{1,1}^1, \dots, \tilde{\mathbf{Y}}_{b,t}^J = \mathbf{Z}\mathbf{A}_{1,1}^{1\text{Test}} \otimes \mathbf{SBRMY}_{1,1}^1, \dots, \mathbf{Z}\mathbf{A}_{b,t}^{J\text{Test}} \otimes \mathbf{SBRMY}_{b,t}^J \tag{29}$$

Now, we recombine the phase spectrum $\mathbf{Z}\mathbf{P}_{1,1}^1, \dots, \mathbf{Z}\mathbf{P}_{b,t}^J$ with the estimated source signal magnitude spectra $\tilde{\mathbf{X}}_{1,1}^1, \dots, \tilde{\mathbf{X}}_{b,t}^J$ and $\tilde{\mathbf{Y}}_{1,1}^1, \dots, \tilde{\mathbf{Y}}_{b,t}^J$ to obtain the modified complex spectra $\tilde{\mathbf{X}}_{1,1}^1(\tau, f), \dots, \tilde{\mathbf{X}}_{b,t}^J(\tau, f)$ and $\tilde{\mathbf{Y}}_{1,1}^1(\tau, f), \dots, \tilde{\mathbf{Y}}_{b,t}^J(\tau, f)$ using Eqs. (30) and (31), respectively.

$$\tilde{\mathbf{X}}_{1,1}^1(\tau, f), \dots, \tilde{\mathbf{X}}_{b,t}^J(\tau, f) = \tilde{\mathbf{X}}_{1,1}^1 e^{i\mathbf{Z}\mathbf{P}_{1,1}^1}, \dots, \tilde{\mathbf{X}}_{b,t}^J e^{i\mathbf{Z}\mathbf{P}_{b,t}^J} \tag{30}$$

$$\tilde{\mathbf{Y}}_{1,1}^1(\tau, f), \dots, \tilde{\mathbf{Y}}_{b,t}^J(\tau, f) = \tilde{\mathbf{Y}}_{1,1}^1 e^{i\mathbf{Z}\mathbf{P}_{1,1}^1}, \dots, \tilde{\mathbf{Y}}_{b,t}^J e^{i\mathbf{Z}\mathbf{P}_{b,t}^J} \tag{31}$$

The ISTFT is used to convert the modified complex source signal spectra $\tilde{\mathbf{X}}_{1,1}^1(\tau, f), \dots, \tilde{\mathbf{X}}_{b,t}^J(\tau, f)$ and $\tilde{\mathbf{Y}}_{1,1}^1(\tau, f), \dots, \tilde{\mathbf{Y}}_{b,t}^J(\tau, f)$ into the modified subband signals $\tilde{\mathbf{x}}_{1,1}^1, \dots, \tilde{\mathbf{x}}_{b,t}^J$ and $\tilde{\mathbf{y}}_{1,1}^1, \dots, \tilde{\mathbf{y}}_{b,t}^J$. Finally, the estimated source signals $\tilde{\mathbf{x}}(t)$ and $\tilde{\mathbf{y}}(t)$ are obtained by applying the IDTCWT to the subband signals $\tilde{\mathbf{x}}_{1,1}^1, \dots, \tilde{\mathbf{x}}_{b,t}^J$ and $\tilde{\mathbf{y}}_{1,1}^1, \dots, \tilde{\mathbf{y}}_{b,t}^J$. The proposed algorithm for the training and testing stages is shown in Table 2.

6 Evaluations and Results

To evaluate the effectiveness of the proposed algorithm, we compare it with the STFT – SNMF [11, 12, 33, 37, 39, 41] and JDL [34] models. In these simulations, we use

Table 2 Proposed algorithms for the training and testing stages**Training Algorithm:**

Input: Training sets $\mathbf{x}(t)$ and $\mathbf{y}(t)$, decomposition level J , subband index b , tree level tl , window length, basis vector size, number of iterations, and sparsity constant μ

Output: $\mathbf{XYW}_{1,1}^1, \dots, \mathbf{XYW}_{b,tl}^J$

Step 1: Calculate the wavelet coefficients $\mathbf{x}_{1,1}^1, \dots, \mathbf{x}_{b,tl}^J$ and $\mathbf{y}_{1,1}^1, \dots, \mathbf{y}_{b,tl}^J$ via the DTCWT

Step 2: Obtain the complex spectra $\mathbf{X}_{1,1}^1(\tau, f), \dots, \mathbf{X}_{b,tl}^J(\tau, f)$ and $\mathbf{Y}_{1,1}^1(\tau, f), \dots, \mathbf{Y}_{b,tl}^J(\tau, f)$ by applying the STFT

Step 3: Obtain the magnitude spectra $\mathbf{XA}_{1,1}^1, \dots, \mathbf{XA}_{b,tl}^J$ and $\mathbf{YA}_{1,1}^1, \dots, \mathbf{YA}_{b,tl}^J$ by taking the absolute values of the complex spectra

Step 4: Determine the basis matrices $\mathbf{XW}_{1,1}^1, \dots, \mathbf{XW}_{b,tl}^J$ and $\mathbf{YW}_{1,1}^1, \dots, \mathbf{YW}_{b,tl}^J$ in accordance with Eqs. (21) and (22)

Step 5: Concatenate these basis matrices as follows: $\mathbf{XYW}_{1,1}^1, \dots, \mathbf{XYW}_{b,tl}^J = \left[\mathbf{XW}_{1,1}^1 \ \mathbf{YW}_{1,1}^1 \right], \dots, \left[\mathbf{XW}_{b,tl}^J \ \mathbf{YW}_{b,tl}^J \right]$

Testing Algorithm:

Input: Mixed signal $\mathbf{z}(t)$, concatenated basis matrices $\mathbf{XYW}_{1,1}^1, \dots, \mathbf{XYW}_{b,tl}^J$ learned in the training stage, decomposition level J , subband index b , tree level tl , and the number of iterations

Output: Estimated separate signals $\tilde{\mathbf{x}}(t)$ and $\tilde{\mathbf{y}}(t)$

Step 1: Compute the wavelet coefficients $\mathbf{z}_{1,1}^1, \dots, \mathbf{z}_{b,tl}^J$ via the DTCWT

Step 2: Acquire the complex spectrum $\mathbf{Z}_{1,1}^1(\tau, f), \dots, \mathbf{Z}_{b,tl}^J(\tau, f)$ by applying the STFT and obtain the magnitude spectrum $\mathbf{ZA}_{1,1}^1, \dots, \mathbf{ZA}_{b,tl}^J$ by taking the absolute values of the complex spectral components

Step 3: Obtain the weight matrices $\mathbf{XH}_{1,1}^1, \dots, \mathbf{XH}_{b,tl}^J$ and $\mathbf{YH}_{1,1}^1, \dots, \mathbf{YH}_{b,tl}^J$ in accordance with Eq. (23)

Step 4: Calculate the initial magnitudes $\bar{\mathbf{X}}_{1,1}^1, \dots, \bar{\mathbf{X}}_{b,tl}^J$ and $\bar{\mathbf{Y}}_{1,1}^1, \dots, \bar{\mathbf{Y}}_{b,tl}^J$ for the source signals using Eqs. (24) and (25)

Step 5: Calculate the subband binary ratio masks $\mathbf{SBRMX}_{1,1}^1, \dots, \mathbf{SBRMX}_{b,tl}^J$ and $\mathbf{SBRMY}_{1,1}^1, \dots, \mathbf{SBRMY}_{b,tl}^J$ in accordance with Eqs. (26) and (27)

Step 7: Estimate the magnitudes $\tilde{\mathbf{X}}_{1,1}^1, \dots, \tilde{\mathbf{X}}_{b,tl}^J$ and $\tilde{\mathbf{Y}}_{1,1}^1, \dots, \tilde{\mathbf{Y}}_{b,tl}^J$ for the source signals using Eqs. (28) and (29)

Step 8: Determine the estimated complex spectra $\tilde{\mathbf{X}}_{1,1}^1(\tau, f), \dots, \tilde{\mathbf{X}}_{b,tl}^J(\tau, f)$ and $\tilde{\mathbf{Y}}_{1,1}^1(\tau, f), \dots, \tilde{\mathbf{Y}}_{b,tl}^J(\tau, f)$ for the source signals using Eqs. (30) and (31)

Step 9: Compute the modified subband signals $\tilde{\mathbf{x}}_{1,1}^1, \dots, \tilde{\mathbf{x}}_{b,tl}^J$ and $\tilde{\mathbf{y}}_{1,1}^1, \dots, \tilde{\mathbf{y}}_{b,tl}^J$ via the ISTFT

Step 10: Obtain the estimated source signals $\tilde{\mathbf{x}}(t)$ and $\tilde{\mathbf{y}}(t)$ by applying the IDTCWT

speech signals from the GRID audiovisual corpus [2], as the training and testing data (including different male and female speech samples). There are 34 speakers (18 male and 16 female speakers), and each speaker speaks 1000 utterances. We concatenate all the utterances together for each speaker. For each speaker, we randomly choose 500 utterances for training and 200 utterances for testing. In these simulations, we use two types of speech signal combinations: one for same-gender (male–male or female–female) speech separation, where the combined signals are denoted by M1 and M2 or by F1 and F2, and another for opposite-gender (male–female) speech

separation, where the combined signals are denoted by M and F. For same-gender signal separation, eight utterances from same-gender speakers' are utilized to form one experimental group, and eight different utterances from same-gender speakers' are used to form another experimental group. For opposite-gender speech separation, we choose sixteen male speakers to compose one experimental group and sixteen female speakers to compose another experimental group. The length of each training signal is approximately 60 s, and the length of each testing signal is approximately 10 s. The sampling rate for each speech signal is 8000 Hz, and the signal is transformed into the time–frequency domain by using a 512-point STFT with 50% overlap.

The speech separation performance is evaluated in terms of the signal-to-interference ratio (SIR) [10], source distortion ratio (SDR) [10], average frequency-weighted segmental SNR (fwsegSNR) [38], short-time objective intelligibility (STOI) [35], perceptual evaluation of speech quality (PESQ) [27], Hearing-Aid Speech Perception Index (HASPI) [15], and Hearing-Aid Speech Quality Index (HASQI) [16]. The SDR value estimates the overall speech quality; it is defined as the ratio of the power of the input signal to the power of the difference between the input and reconstructed signals. Higher SDR scores indicate better performance. In addition to the SDR, the SIR captures the error caused by failure to remove interfering signal information during the SS procedure. A higher SIR indicates higher separation quality. The STOI is defined as the correlation between the short-term temporal envelopes of the clean and separated speech signals, and its value ranges from 0 to 1, with a higher STOI score indicating better intelligibility. We choose the fwsegSNR as the objective measure to evaluate the intelligibility of the captured speech signal; a higher value represents better performance. For both hearing-impaired patients and people with normal hearing, the HASQI and HASPI gauge sound quality and perception, respectively. Similar to the STOI, the values range from 0 to 1, and higher scores indicate better sound quality and intelligibility, respectively.

First, the separation results of the different strategies are shown in Fig. 6, where the original female and male speech spectrograms are displayed in Fig. 6a and b, respectively. The estimated male speech spectrograms are presented in Fig. 6c, e, and g for the DWT – STFT – SNMF algorithm, the SWT – SNMF algorithms, and the proposed method (PM), respectively, and the female speech spectrograms are similarly presented in Fig. 6d, f, and h. From this figure, we can see that the SS quality of the DWT – STFT – SNMF method is poor due to the total rejection of the high-frequency component, where the SWT – SNMF method adds unwanted speech components to the estimated male and female speech signals. By contrast, the PM recovers male and female speech signals that are approximately similar to the original signals.

Second, in Fig. 7, we compare the PM with the existing models in terms of the fwsegSNR. From this figure, it appears that the PM performs very well in all cases compared to the other current methods. In the various SS scenarios, our method improves the fwsegSNR scores by 20.33% for the M1 signal, 17.76% for the M2 signal, 24.93% for the F1 signal, 32.17% for the F2 signal, 30.49% for the M signal, and 15.70% for the F signal compared to the DTCWT – SNMF method.

Third, in Fig. 8, we show that in terms of the SDR and SIR, the PM considerably outperforms the existing models, namely the STFT – SNMF, DWT – STFT – SNMF, DWPT – SNMF, SWT – SNMF, DTCWT – SNMF, CJD, and JDL algorithms. For

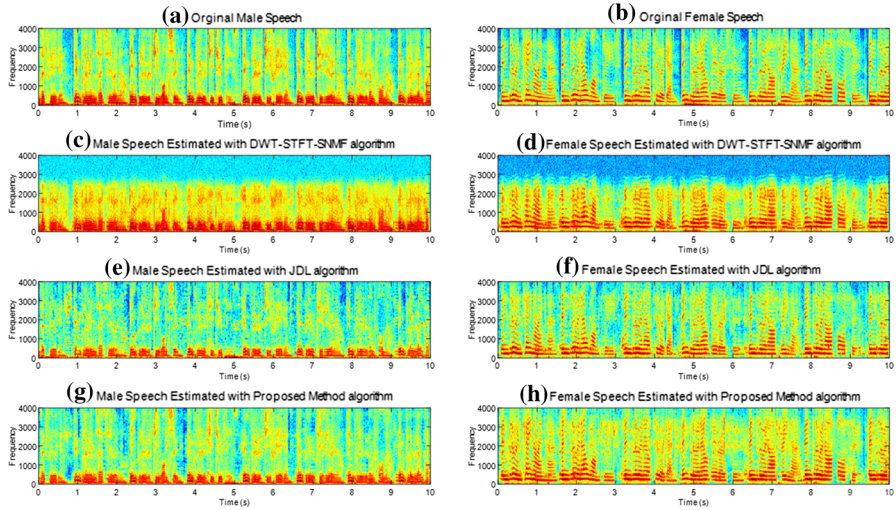


Fig. 6 Spectrograms of the original male and female speech signals, and the male and female speech signals, recovered with the DWT–STFT–SNMF algorithm, the SWT–SNMF algorithm, and the PM, where the x-axis corresponds to the time in seconds, and the y-axis corresponds to the frequency in kHz

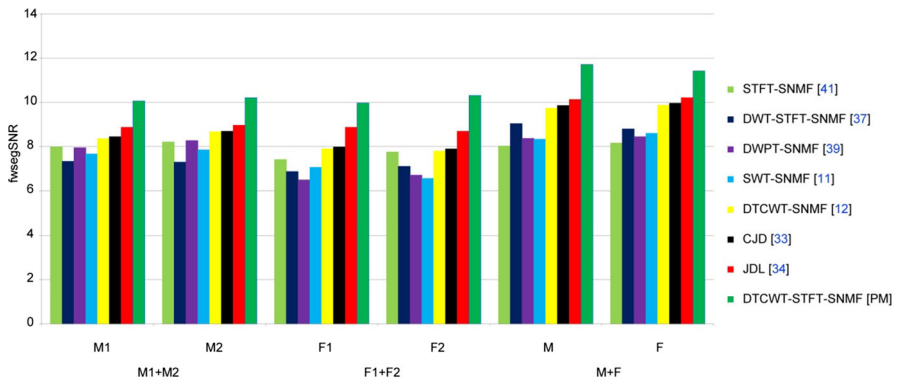


Fig. 7 Comparison of the fwsegSNR values of the eight methods for the same- and opposite-gender cases

all cases of speech separation, the SDR values of the PM are higher than those of the existing models. With the PM, the SDR is improved from 4.72 to 5.29 dB for the M1 signal, from 4.39 to 5.24 dB for the M2 signal, from 5.27 to 5.97 dB for the F1 signal, from 4.61 to 5.85 dB for the F2 signal, from 7.65 to 9.01 dB for the M signal, and from 6.35 to 9.12 dB for the F signal compared to the DWT – STFT – SNMF model. From this figure, we can also see that the SIR values of the estimated signals are better with the PM than with the existing models. Moreover, we find that the separation result of the opposite-gender signals is much better than those for the same-gender signals.

Fourth, Tables 3 and 4 present a comparative performance analysis of the PM and the existing methods in terms of the STOI and PESQ. Compared to the existing DWT – STFT – SNMF method, the PM improves the STOI scores by 13.24% and

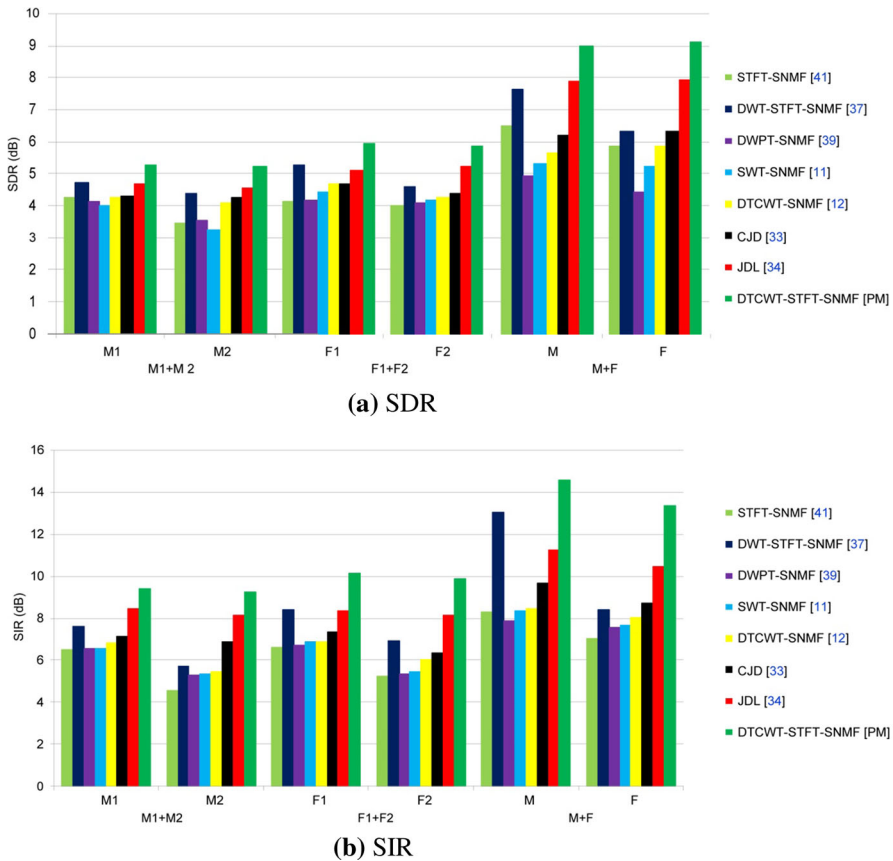


Fig. 8 Comparative performance evaluation of the existing models and the PM using: (a) the SDR and (b) the SIR for the same- and opposite-gender cases

10.98% for the M1 and M2 signals, respectively; by 13.20% and 11.73% for the F1 and F2 signals, respectively; and by 24.96% and 10.26% for the M and F signals, respectively. From Table 4, we can also see that the PESQ scores of the estimated signals are better than with the existing models.

Fifth, Tables 5 and 6 present the HASQI and HASPI results of the different methods, namely the STFT – SNMF, DWT – STFT – SNMF, DWPT – SNMF, SWT – SNMF, DTCWT – SNMF, CJD, and JDL methods and the PM for the same- and opposite-gender speech separation tasks. From Table 5, one can observe that the PM algorithm yields better HASPI values than the other algorithms for all cases of separation. The PM algorithm outperforms the other seven methods in terms of the HASQI results for all cases of separation.

Sixth, we present the SS performance achieved on the TIMIT database [6] to further confirm the superiority of the PM in mixed speech separation experiments. For these experiments, 24 speakers (12 male and 12 female speakers) were selected from the TIMIT database. Each speaker utters ten sentences, corresponding to a total of

Table 3 Comparison of the STOI values achieved with the eight methods for the same- and opposite-gender cases

Case Method	M1 + M2		F1 + F2		M + F	
	M1	M2	F1	F2	M	F
STFT–SNMF [41]	0.717	0.763	0.743	0.737	0.776	0.747
DWT–STFT–SNMF [37]	0.702	0.710	0.689	0.699	0.705	0.760
DWPT–SNMF [39]	0.719	0.724	0.674	0.744	0.745	0.702
SWT–SNMF [11]	0.723	0.727	0.710	0.754	0.769	0.714
DTCWT–SNMF [12]	0.725	0.730	0.730	0.764	0.779	0.744
CJD [33]	0.726	0.738	0.735	0.767	0.781	0.748
JDL [34]	0.746	0.768	0.785	0.787	0.793	0.778
DTCWT–STFT–SNMF [PM]	0.795	0.788	0.780	0.781	0.881	0.838

Bold values indicate best result

Table 4 Comparison of the PESQ values achieved with the eight methods for the same- and opposite-gender cases

Case Method	M1 + M2		F1 + F2		M + F	
	M1	M2	F1	F2	M	F
STFT–SNMF [41]	2.001	2.101	2.012	2.006	2.217	2.212
DWT–STFT–SNMF [37]	2.116	2.128	2.086	2.069	2.386	2.263
DWPT–SNMF [39]	2.009	2.111	2.016	2.011	2.223	2.225
SWT–SNMF [11]	2.011	2.116	2.018	2.015	2.238	2.229
DTCWT–SNMF [12]	2.012	2.121	2.035	2.028	2.244	2.234
CJD [33]	2.023	2.037	2.045	2.039	2.254	2.253
JDL [34]	2.067	2.073	2.072	2.062	2.321	2.331
DTCWT–STFT–SNMF [PM]	2.117	2.154	2.131	2.144	2.464	2.374

Bold values indicate best result

Table 5 Comparison of the HASPI values achieved with the eight methods for the same- and opposite-gender cases

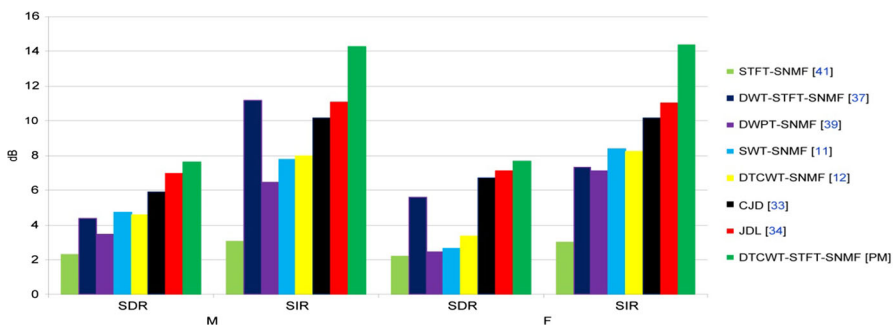
Case Method	M1+M2		F1+F2		M+F	
	M1	M2	F1	F2	M	F
STFT–SNMF [41]	0.9942	0.9956	0.9954	0.9956	0.9982	0.9987
DWT–STFT–SNMF [37]	0.9724	0.9573	0.9761	0.9576	0.9951	0.9915
DWPT–SNMF [39]	0.9731	0.9891	0.9831	0.9770	0.9888	0.9850
SWT–SNMF [11]	0.9767	0.9900	0.9845	0.9854	0.9939	0.7514
DTCWT–SNMF [12]	0.9785	0.9911	0.9857	0.9817	0.9917	0.9894
CJD [33]	0.9793	0.9944	0.9863	0.9845	0.9943	0.9869
JDL [34]	0.9813	0.9947	0.9896	0.9885	0.9964	0.9878
DTCWT–STFT–SNMF [PM]	0.9961	0.9967	0.9971	0.9964	0.9995	0.9991

Bold values indicate best result

Table 6 Comparison of the HASQI values achieved with the eight methods for the same- and opposite-gender cases

Case	M1+M2		F1+F2		M+F	
	M1	M2	F1	F2	M	F
STFT–SNMF [41]	0.412	0.405	0.407	0.414	0.555	0.503
DWT–STFT–SNMF [37]	0.269	0.246	0.271	0.257	0.485	0.445
DWPT–SNMF [39]	0.321	0.352	0.313	0.404	0.450	0.430
SWT–SNMF [11]	0.322	0.352	0.328	0.428	0.488	0.418
DTCWT–SNMF [12]	0.333	0.364	0.330	0.420	0.470	0.452
CJD [33]	0.346	0.373	0.349	0.421	0.483	0.461
JDL [34]	0.389	0.398	0.401	0.429	0.521	0.493
DTCWT–STFT–SNMF [PM]	0.439	0.437	0.443	0.447	0.609	0.569

Bold values indicate best result

**Fig. 9** Comparative performance evaluation of the existing and the PM in terms of the SDR and SIR for the opposite-gender case

240 sentences. Of the 10 sentences uttered by each distinct speaker, the first eight sentences are selected for training, and the remaining two sentences are used for testing. To investigate the performance of our proposed strategy, we consider the SDR, SIR, STOI, and PESQ scores. From Fig. 9 and Table 5, one can clearly see that the proposed strategy performs better in strengthening speech than the other seven techniques (STFT – SNMF, DWT – STFT – SNMF, DWPT – SNMF, SWT – SNMF, DTCWT – SNMF, CJD, and JDL) according to the SDR, SIR, STOI, and PESQ scores for opposite-gender signal separation.

Finally, for a practical comparison of the computational load, we compare the execution times required to generate an estimated signal during one iteration when the analysis is implemented in MATLAB on a PC equipped with an Intel® Core™ i7-4790 CPU @ 3.60 GHz. As seen from the results in Table 8, the execution time of the PM is shorter than that those of all other methods except the STFT – SNMF and DWT – STFT – SNMF algorithms, while the other metrics of the PM (see Tables 3, 4, 5, 6, 7 and Figs. 7, 8, 9) are better than those of the other methods. The P/O ratios (where P refers to the PM and O signifies another method considered for comparison)

Table 7 Comparison of the PESQ and STOI values achieved with the eight methods for the opposite-gender case

Method	PESQ		STOI	
	M	F	M	F
STFT–SNMF [41]	2.204	1.097	0.757	0.439
DWT–STFT–SNMF [37]	2.376	2.170	0.708	0.673
DWPT–SNMF [39]	2.273	1.898	0.763	0.700
SWT–SNMF [11]	2.356	1.983	0.787	0.721
DTCWT–SNMF [12]	2.335	1.938	0.780	0.717
CJD [33]	2.324	2.012	0.789	0.719
JDL [34]	2.375	2.141	0.805	0.730
DTCWT–STFT–SNMF [PM]	2.509	2.192	0.835	0.770

Bold values indicate best result

Table 8 Comparison of the computational loads

Method	Method Execution time (Seconds)	Ratio (P/O)
STFT–SNMF [41]	0.0795	2.3
DWT–STFT–SNMF [37]	0.148	1.24
DWPT–SNMF [39]	0.673	0.272
SWT–SNMF [11]	0.249	0.735
DTCWT–SNMF [12]	0.198	0.93
CJD [33]	0.798	0.229
JDL [34]	0.721	0.254
DTCWT–STFT–SNMF [PM]	0.183	–

in terms of the execution times are also listed in Table 8. The results confirm that the proposed method can reduce the online computational load by factors of approximately 3.68, 1.36, 1.08, 4.36, and 3.94, relative to the conventional methods listed in the table sequentially from top to bottom, respectively, except for the STFT – SNMF and DWT – STFT – SNMF algorithms.

7 Conclusion

In this paper, we have proposed an improved speech separation model using the DTCWT and the STFT with SNMF. The development of this dual-transform (DTCWT and STFT)-based speech separation model is the main focus of our research. First, we apply the DTCWT and STFT successively to the input speech signal to provide a more flexible basic framework for improved feature extraction. Second, the speech signal is sparsely represented by applying SNMF to obtain the corresponding weight matrices considering only the magnitude spectrograms used in the testing phase. Finally, the estimated separated speech signals are generated by applying the ISTFT and IDTCWT consecutively. The DTCWT separates the high- and low-frequency components of the

time-domain signal by means of filters, and the STFT accurately characterizes the time–frequency components. For this reason, the quality and intelligibility of the separated speech signals are improved compared with the results of existing methods. In evaluations of the improvement in the separated speech signals using various evaluation methods, the experimental outcomes demonstrate that the overall performance of the proposed speech separation model is superior to that of the existing models. In the future, we plan to investigate alternative training and testing algorithms using deep neural networks.

Acknowledgements This research was supported by the National Natural Science Foundation of China (Grant No. 61671418).

Data availability The datasets generated or analyzed during the current study are not publicly available because they are the subject of ongoing research but are available from the first author upon reasonable request.

References

1. G. Bao, Y. Xu, Z. Ye, Learning a discriminative dictionary for single-channel speech separation. *IEEE Trans. Audio Speech Lang. Process.* **22**(7), 1130–1138 (2014)
2. M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **120**(5), 2421 (2006)
3. D.L. Daniel, H.S. Seung, Learning the pans of objects with non-negative matrix factorization. *Nature* **401**, 788–791 (1999)
4. M.G. Emad, E. Hakan, Single channel speech music separation using nonnegative matrix factorization with sliding windows and spectral masks. *Digital Signal Processing (DSP)*, in *17th International Conference in August* (2011)
5. G.G. Francois, J.M. Gautham, Stopping criteria for non-negative matrix factorization based supervised and semi-supervised source separation. *IEEE Signal Processing Letters*, November (2014)
6. J. Garofolo, et al., TIMIT acoustic-phonetic continuous speech corpus. LDC93S1 (1993)
7. E.M. Grais, H. Erdogan, Discriminative non-negative dictionary learning using cross-coherence penalties for single channel source separation, in *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)*. Lyon, France, 25–29 August (2013)
8. R. Hidayat, A. Bejo, S. Sumaryono, A. Winursito, Denoising speech for MFCC feature extraction using wavelet transformation in speech recognition system, in *10th International Conference on Information Technology and Electrical Engineering* (2018)
9. P.O. Hoyer, Non-negative matrix factorization with sparseness constraint. *J. Mach. Learn. Res.* 1457–1469, November (2004)
10. Y. Hu, P.C. Loizou, Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **16**(1), 229–238 (2008)
11. M.S. Islam, T.H. Al Mahmud, W.U. Khan, Z. Ye, Supervised single-channel speech enhancement based on stationary wavelet transforms and non-negative matrix factorization with concatenated framing process and subband smooth ratio mask. *J. Sig. Process. Syst. Signal. Image Video Technol.* 1–14 (2019)
12. M.S. Islam, T.H. Al Mahmud, W.U. Khan, Z. Ye, 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**, 353 (2019)
13. G.J. Jang, T.W. Lee, A maximum likelihood approach to single-channel source separation. *J. Mach. Learn. Res.* **4**, 1365–1392 (2003)
14. D.S. Kapoor, A.K. Kohli, Gain adapted optimum mixture estimation scheme for single-channel speech separation. *Circuits Syst. Signal Process.* **32**(5), 2335–2351 (2013)
15. J.M. Kates, K.H. Arehart, The hearing-aid speech perception index (HASPI). *Speech Commun.* **65**, 75–93 (2014)

16. J.M. Kates, K.H. Arehart, The hearing-aid speech quality index (HASQI). *J. Audio Eng. Soc.* **58**, 5363–5381 (2010)
17. N.G. Kingsbury, 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. 8–11 Sept (1998)
18. R.J. Le, F.J. Weninger, J.R. Hershey, Sparse NMF half-baked or well done? technical report TR2015–023, Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA, March (2015)
19. D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* **13**, 556–562 (2001)
20. A. Mahmoodzadeh, H.R. Abutaleb, Hybrid approach to single-channel speech separation based on coherent incoherent modulation filtering. *Circuits Syst. Signal Process.* **36**(5), 1970–1988 (2017)
21. S. Mavaddati, A novel singing voice separation method based on sparse non-negative matrix factorization and low-rank modeling. *Iran. J. Electr. Electron. Eng.* **15**, 2 (2019)
22. P. Mercorelli, A denoising procedure using wavelet packets for instantaneous detection of pantograph oscillations. *Mech. Syst. Signal Process.* **35**, 137–149 (2013)
23. P. Paatero, U. Tapper, Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**(2), 111–126 (1994)
24. B.A. Pearlmutter, R.K. Olsson, Linear program differentiation for single-channel speech separation, in *16th IEEE Signal Processing Society Workshop in MLSP, Arlington, VA, USA* (2006)
25. T. Pham, Y.S. Lee, Y.B. Lin, T.C. Tai, J.C. Wang, Single channel source separation using sparse nmf and graph regularization. *ASE Big Data Soc. Inform.* **55**, 1–7 (2015)
26. B. Premanode, J. Vongprasert, C. Toumazou, Noise reduction for nonlinear nonstationary time series data using averaging intrinsic mode function. *Algorithms* **6**(3), 407–429 (2013)
27. A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in *IEEE International Conference on Acoustics, Speech, Signal Processing*. 6, 7–11 May (2001)
28. S.T. Roweis, One microphone source separation. *Advances in Neural Information Processing Systems*. 793–799 (2001).
29. S.T. Roweis, Factorial models and refiltering for speech separation and denoising, in *Eurospeech*, Geneva, 1009–1012 (2003)
30. M.N. Schmidt, R.K. Olsson, Single-channel speech separation using sparse non-negative matrix factorization, in *9th International Conference on Spoken Language Processing*. Pittsburgh, PA, USA (2006)
31. M.N. Schmidt, M. Morup, Sparse non-negative matrix factor 2-D deconvolution for blind single-channel source separation. *Indep. Compon. Anal. Blind Signal Sep.* **3889**, 700–707 (2006)
32. S.M. Seedahmed, A generalised wavelet packet-based anonymization approach for ECG security application. *9, 18*, 6137–6147 (2016)
33. L. Sun, C. Zhao, M. Su, F. Wang, Single-channel blind source separation based on joint dictionary with common sub-dictionary. *Int. J. Speech Technol.* **21**(1), 19–27 (2018)
34. L. Sun, K. Xie, T. Gu, J. Chen, Z. Yang, Joint dictionary learning using a new optimization method for single-channel blind source separation. *Speech Commun.* **106**, 85–94 (2019)
35. C.H. Tall, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **19**(7), 2125–2136 (2011)
36. P. Tianliang, C. Yang, L. Zengli, A time-frequency domain blind source separation method for underdetermined instantaneous mixtures. *Circuits Syst. Signal Process.* **34**(12), 3883–3895 (2015)
37. Y.V. Varshney, Z.A. Abbasi, M.R. Abidi, O. Farooq, Frequency selection based separation of speech signals with reduced computational time using sparse NMF. *Arch. Acoust.* **42**(2), 287–295 (2017)
38. E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)
39. S. Wang, A. Chern, Y. Tsao, J. Hung, X. Lu, Y. Lai, B. Su, Wavelet speech enhancement based on non-negative matrix factorization. *IEEE Signal Process. Lett.* **23**, 1101–1105 (2016)
40. Y. Wang, Y. Li, K.C. Ho, A. Zare, M. Skubic, Sparsity promoted non-negative matrix factorization for source separation and detection, in *Proceedings of the 19th International Conference on Digital Signal Processing*. IEEE. 20–23 August (2014).
41. Z. Wang, F. Sha, Discriminative non-negative matrix factorization for single-channel speech separation, in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2014)

42. Y. Xu, G. Bao, X. Xu, Z. Ye, Single-channel speech separation using sequential discriminative dictionary learning. *Signal Process.* **106**, 134–140 (2015)
43. V.V. Yash, A.A. Zia, R.A. Musiur, O. Farooq, Variable sparsity regularization factor based SNMF for monaural speech separation, in *40th International Conference on Telecommunications and Signal Processing (TSP)*. 5–7 July (2017)

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