Annals of Emerging Technologies in Computing (AETiC)
Vol. 8, No. 1, 2024

Research Article

Single-channel Speech Separation Based on Double-density Dual-tree CWT and SNMF

Md Imran Hossain, Md Abdur Rahim* and Md Najmul Hossain

Pabna University of Science and Technology, Pabna, Bangladesh
imran05ice@pust.ac.bd; rahim@pust.ac.bd; najmul_eece@pust.ac.bd
*Correspondence: rahim@pust.ac.bd

Received: 8th June 2023; Accepted: 25th December 2023; Published: 1st January 2024

Abstract: Speech is essential to human communication; therefore, distinguishing it from noise is crucial. Speech separation becomes challenging in real-world circumstances with background noise and overlapping speech. Moreover, the speech separation using short-term Fourier transform (STFT) and discrete wavelet transform (DWT) addresses time and frequency resolution and time-variation issues, respectively. To solve the above issues, a new speech separation technique is presented based on the double-density dual-tree complex wavelet transform (DDDTCWT) and sparse non-negative matrix factorization (SNMF). The signal is separated into high-pass and low-pass frequency components using DDDTCWT wavelet decomposition. For this analysis, we only considered the low-pass frequency components and zeroed out the high-pass ones. Subsequently, the STFT is then applied to each sub-band signal to generate a complex spectrogram. Therefore, we have used SNMF to factorize the joint form of magnitude and the absolute value of real and imaginary (RI) components that decompose the basis and weight matrices. Most researchers enhance the magnitude spectra only, ignore the phase spectra, and estimate the separated speech using noisy phase. As a result, some noise components are present in the estimated speech results. We are dealing with the signal's magnitude as well as the RI components and estimating the phase of the RI parts. Finally, separated speech signals can be achieved using the inverse STFT (ISTFT) and the inverse DDDTCWT (IDDDTCWT). Separation performance is improved for estimating the phase component and the shift-invariant, better direction selectivity, and scheme freedom properties of DDDTCWT. The speech separation efficiency of the proposed algorithm outperforms performance by 6.53–8.17 dB SDR gain, 7.37-9.87 dB SAR gain, and 14.92–17.21 dB SIR gain compared to the NMF method with masking on the TIMIT dataset.

Keywords: ***Double Density Dual-Tree Complex Wavelet Transform; Speech Separation; Sparse Non-negative Matrix Factorization; Short-time Fourier Transform***

1. Introduction

In recent years, single-channel speech separation has drawn significant scientific interest due to the continuously expanding number of voice-based solutions for real-world applications. A human can recognize individual speech from an interference signal, even in complicated and chaotic acoustical environments. Speech separation methods with better accuracy have not yet been developed. The main goal of the present digital world is to solve all problems digitally. That's why speech separation has attracted remarkable attention from researchers. The aim of speech separation (SS) is to estimate a target speech from noisy signals. It has many potential usages in real-life applications such as assisted living systems, hearing aid devices, local police investigations on recorded speech, automatic speech recognition (ASR), teleconferencing systems, and controlling humanoid robots [1, 2]. Depending on the variety of channels, speech separation concerns are categorized as single-channel, multichannel, or binaural. A single-channel SS (SCSS) method [3–5] is complicated because only a single recording is obtainable, and the description that may be retrieved is limited.

Md Imran Hossain, Md Abdur Rahim and Md Najmul Hossain, "Single-channel Speech Separation Based on Double-density Dual-tree CWT and SNMF", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 1-12, Vol. 8, No. 1, 1st January 2024, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2024.01.001, Available: http://aetic.theiaer.org/archive/v8/v8n1/p1.html.

Most SCSS approaches can be divided into two types: those based on computational auditory scene analysis (CASA) and those based on models. The CASA-based algorithm separates speech by considering the human hearing mechanism [6]. Model-based SCSS algorithms mainly rely on the training data of the source signal. The training data models were generated using probabilistic models, specifically the Gaussian mixed model (GMM) [7] and the hidden Markov model (HMM) [8]. These representations are generally used in source separation processes. The supposition is that during the training and separation stages, the energy level of a mixed signal is equivalent to the energy level of source signals. Non-negative matrix factorization is an alternative training data model [9]. In both the training and separation phases, these models place no constraints on the magnitude of the energy disparity between the source signals. However, the number of speech signals is limited. In [10], the authors used sparse non-negative matrix factorization (SNMF). They learned the sparse representation of the data using SNMF to address the challenge of differentiating many speech sources from a single microphone recording.

Most of these methods use the short-term Fourier transform STFT domain [11]. The STFT examines a time-domain signal in tiny segments or frames to evaluate whether it is stationary. It necessitates using a window function, which is significant in such circumstances. Despite the weak frequency resolution, we obtained a better time resolution and more stationery through a narrow window selection. Furthermore, selecting a wider window gives better frequency resolution and worse static estimates; however, time resolution is poor. Furthermore, STFT suffers from this time-frequency difficulty due to a lack of knowledge about which frequency exists at which moments. Because we do not know what frequency exists at what time instance, the STFT suffers from this time-frequency resolution problem. WT is a useful tool for modeling and examining non-stationary signals; for example, speech signals. These signals demonstrate slow temporal variations in low frequencies and sudden changes in high frequencies.

Recently, wavelet-based separation methods [12-14] have emerged for researchers to overcome the abovementioned problems. In [12], the discrete wavelet transform (DWT) divides a signal into low-frequency approximation and high-frequency details coefficients. This method has reduced separation time, but the separated sign seriously affects individual speakers' intelligibility. However, this method has redundancy difficulties and cannot take advantage of the sparseness of distinct speech signals. To use both transformations and get a higher level of resolution when processing the mixture, we presented an SS approach based on the DTCWT and STFT [14]. To overwhelm the above-revealed problems, we recommend an SCSS method by considering the properties of speech signals. To overwhelm the above-revealed problems, we recommend an SCSS method by considering the sequential use of DDDTCWT and STFT makes the signal more stationary, resulting in a better transformation.

We suggest an SCSS technique that uses the dual-domain transform and sparse representation to overcome the aforementioned issues. The following is a synopsis of the paper's contribution:

1. Many other approaches merely enhance the magnitude spectra while ignoring the phase spectra and estimating the separated speech using noisy phases. These techniques do not fully exploit all the information contained in the waveforms of the signals. We calculate the phase spectrum from the real and imaginary (RI) components and then mix it with the magnitude spectrum in our proposed strategy. The enhanced phase increases separation performance while decreasing noise, artifacts, and interference.
2. We enhanced only the low-pass frequency components; the high-pass frequency signals were set to zero. Due to ignoring the high-frequency components, the model's time complexity is reduced. For approximate shift-invariant, better direction selectivity, and perfect reconstruction properties, the use of DDDTCWT improves the model's separation capability.

2. Literature Review

Humans are remarkably efficient at differentiating speech from mixed or noisy speech by nature. Although they have not yet reached their full potential, researchers are working to develop SS systems that can function similarly to the human auditory system. However, several strategies have been devised to discriminate between single-channel speech signals by taking into account learning techniques, power levels, frequency components, auditory processes, and more.

The authors used a dictionary learning (DL) algorithm to develop the model-based SCSS [15-16]. They believe speech signals with sparse exposure from different speakers have specific distinguishing characteristics. In [15], sequential discriminative dictionary learning (SDDL) measures unique and similar sections of varied voice waveforms. The Discriminative Dictionary Learning (DDL) technique [16] posits that each speaker's speech signal has distinct components. However, the dictionary is constructed from various sources, and there is no guarantee that each atom comes from a single source rather than a mix of them. An over-complete dictionary that allows for sparse signal representation can be built by changing its content to fit a set of signal instances, or it can be selected as a predetermined set of functions. Usually speaking, a joint dictionary, on the other hand, is a redundant dictionary. Although sparse constraints are used to train the dictionaries, one source signal replies to the categorized sub-dictionaries with more sources that cannot be avoided.

The authors of [17] demonstrate the establishment of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for voice recognition and speech separation. They showed the PCA as a great tool for voice recognition, and ICA can separate the signal near about original signal. A singing voice separation method using Robust PCA is presented [3]. The repetition structure of music accompaniment can be regarded as a low-rank subspace, and singing voices can be considered sparse inside the songs. Another ICA-based method is suggested in [18], where the frequency domain representation of the noisy signal is used. This method can reduce the background noise of two-source systems, but it is not relevant where more than two sources exist. These above-mentioned SS algorithms require prior information about the source and mixed signals. Former information about the predicted mean and variance is required for ICA and PCA.

NMF-based algorithms are used iteratively to optimize the cost function [9]. In Bayesian NMF (BNMF), researchers executed numerous preceding structures [19] above the probabilistic model of NMF. In some circumstances, they show an increase in the rate of source detection. Still, extra development in mixed SS is required. Discriminative learning of NMF [20] is used to optimize all basis vectors jointly that reconstruct both clean and mixed-signal. To extract the speeches from the music signal, a SCSS algorithm based on NMF and the combination of cost functions is presented [14]. In the training stage for music, Itakura-Saito (IS) divergence is applied as the NMF cost function, and KL divergence is used as the NMF cost function in the speech training stage. A linear combination of two divergences and a regularization term are used for decomposition. The authors proposed a three-stage hybrid model that can distinguish between two speakers from a single-channel speech mixture in an unsupervised situation [21]. They employed three methods: masking, nonnegative matrix factorization (NMF), and voice segmentation. The authors employed traditional techniques such as NMF, which could lead to overfitting by capturing irrelevant information. However, this paper suggests a novel way to separate speech signals.

In recent years, deep learning has become a popular machine learning technique. In the SS community, deep learning has also received a lot of attention. A deep neural network-based post-processing approach is presented for reconstructing the masked frequency components [22]. They developed a regression from the dependable frequency components to the masked components. After the masked-based isolation, the consistent components are retained unchanged, and the masked components are recovered according to the outcomes of DNN. In [23], the authors present an end-to-end source separation platform that enables us to predict the isolated speech waveform by directly working on the mixture's raw waveform. A large variety of deep learning-based SS methods have been successfully implemented, with excellent results in improving the desired signal from the mixed signal. Furthermore, it is unsuitable for handling limited features, constrained sources inside the cognitive process, and a higher level of computational complexity. To address the mixed speech problem, this paper proposes a new approach based on DDDTCWT and sparse SNMF. The proposed strategy performs better than the earlier approaches described in this work when measured using a variety of objective metrics, including SDR, SIR, and SAR.

3. Proposed Speech Separation (SS) Algorithm

This part represents the recently proposed SCSS approach of the connected substance associated with the proposed technique. Most speech separation algorithms concentrate on the speech signal's STFT, which only examines the magnitude spectrum and ignores the phase spectrum. This paper uses DDDTCWT and

STFT sequentially with a sparse non-negative matrix factorization considering the magnitude, real, and imaginary components. The STFT typically separates a time domain input signal into discrete frames that are individually considered as stationary. However, we don't distinguish what frequency occurs at what instant of existence; the fragment may not be more static. Therefore, we adopt DDDTCWT in our suggested technique, which decomposes the input signal into small segments to extract low- and high-frequency components. Then, for each sub-band signal, STFT is applied to make the signal appear more stationary, leading to a more effective transformation. Finally, the SNMF algorithm is employed to jointly learn the MRI components of the signal after applying DDDTCWT and STFT consecutively. The complete block illustration of the suggested SS algorithm is depicted in detail in Fig. 1. The proposed technique is isolated into two phases: the training phase and the testing phase.

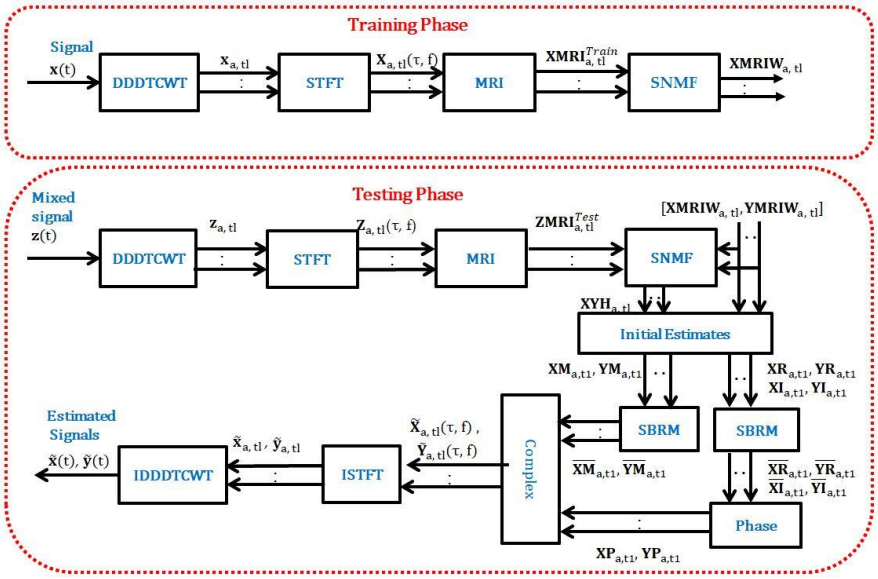

Figure 1. Block diagram of the proposed approach for SS technique

3.1. Double-density Dual-tree Complex Wavelet Transform (DDDTCWT)

Since the Fourier transform does not preserve the time characteristics of a signal; it cannot be used to analyse non-stationary and nonlinear signals. The wavelet transform and its variants are useful for handling non-stationary signals. Double-density dual-tree combined with DWT, each having unique properties and benefits, is known as DDDTCWT [24]. As a result, each of these transforms has its own advantages; however, combining the two creates a superb tool for signal processing applications. Additionally, the double-density DWT provides double the freedom of the system. As a result, we can apply complex and directional wavelet transforms.

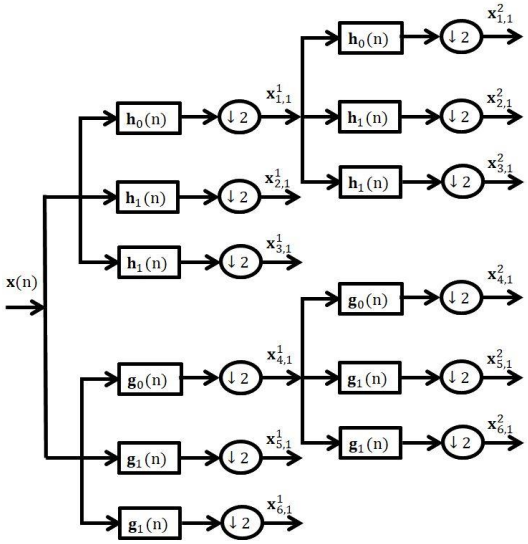

Figure 2. The second level block diagram of DDDTCWT

The DDDTCWT scheme flow is depicted in Fig. 2. We used a two-level decomposition method in this study. It has two distinct filter banks, indicated by the symbols $h_i(n)$ and $g_i(n)$, where $i = 0, 1, 2$.

3.2. Sparse Non-negative Matrix Factorization (SNMF)

The NMF is an investigation algorithm in which the matrix $\mathbf{S} \in \mathbb{R}^{F\times T}$ is disintegrated as a linear product of its bases $\mathbf{W} \in \mathbb{R}^{F\times R}$ and its weights $\mathbf{H} \in \mathbb{R}^{R\times T}$, where the inner dimension R is much less than the matrix S's multiplication of F and T.

$$\mathbf{S} \approx \mathbf{WH} \tag{1}$$

The regulation over sparse output representation can be extended by incorporating sparseness constraints into NMF. When the Euclidean distance and KL divergence cost functions are compared in the sound source partition, the KL cost function shows an outstanding fit [25]. The following is a description of the KL divergence cost function:

$$C_{KL} = \min D(\mathbf{S}||\mathbf{WH}) + \mu||\mathbf{H}||_1 = \sum_{i,j}(\mathbf{S}_{i,j}\log\frac{\mathbf{S}_{i,j}}{(\mathbf{WH})_{i,j}} - \mathbf{S}_{i,j} + (\mathbf{WH})_{i,j}) + \mu\sum_{i,j}|\mathbf{H}_{i,j}| \tag{2}$$

The matrices W and H are expressed by their corresponding Equation (3) and (4), where μ represents the sparsity constant.

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\frac{\mathbf{S}}{\mathbf{WH}}\mathbf{H}^T + \mathbf{W}\otimes\left(\mathbf{1}_v\left(\sum(\mathbf{W}\otimes 1_m\mathbf{H}^T)\right)\right)}{\mathbf{1}_m\mathbf{H}^T + \mathbf{W}\otimes\left(\mathbf{1}_v\left(\sum\left(\mathbf{W}\otimes\left(\frac{\mathbf{S}}{\mathbf{WH}}\mathbf{H}^T\right)\right)\right)\right)} \tag{3}$$

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T\frac{\mathbf{S}}{\mathbf{WH}}}{\mathbf{W}^T 1_m + \mu} \tag{4}$$

Where $\mathbf{1}_m$ is one's matrix, $\mathbf{1}_v$ is a column vector of ones, and all divisions are element-wise.

3.3. Training Phase

We considered two different speech sources delivering signals $\mathbf{x}(t)$ and $\mathbf{y}(t)$ during the training phase. When applied to a speech signal in the time domain, the DDDTCWT separates it into its component high-frequency and low-frequency sub-band signals. Low-frequency components of a signal contain a significant quantity of information, whereas high-frequency components contain almost no information. In this study, we only consider the approximation coefficients corresponding to low-frequency sub-band signals and substitute the detailed coefficients with zero, which are high-frequency sub-band signals to reduce the time complexity. The approximate coefficient is provided by the low pass filter, whereas the high pass filter includes a detail coefficient from the signal $\mathbf{x}(t)$. Each approximate coefficient is subjected to the STFT, which yields the complex forms $\mathbf{X}_{a,tl}(\tau, f)$ that is represented in Equation (5).

$$\mathbf{X}_{a,tl}(\tau, f) = \mathbf{XR}_{a,tl}(\tau, f) + i\,\mathbf{XI}_{a,tl}(\tau, f) \tag{5}$$

Where a, f, and τ indicate the approximation coefficients, frequency, and time bin indices correspondingly. Nowadays, complex spectrums are divided into three parts: the magnitude, the real, and the imaginary. To concatenate the absolute values of the real and imaginary parts with a magnitude part, we use Equation (6).

$$\mathbf{XMRI}_{a,tl}^{Train} = \begin{bmatrix}|\mathbf{XM}_{a,tl}(\tau, f)| \\ |\mathbf{XR}_{a,tl}(\tau, f)| \\ |\mathbf{XI}_{a,tl}(\tau, f)|\end{bmatrix} \tag{6}$$

The combined form of the magnitude spectrum and absolute value of an RI component $\mathbf{XMRI}_{a,tl}^{Train}$ is forwarded to the SNMF. The SNMF decomposes $\mathbf{XMRI}_{a,tl}^{Train}$ into the basis and weight matrices using Equation (7).

$$\mathbf{XMRI}_{a,tl}^{Train} \approx \mathbf{XMRIW}_{a,tl}\mathbf{XH}_{a,tl} + \mu|\mathbf{XH}_{a,tl}|_1 \tag{7}$$

where $\mathbf{XMRIW}_{a,tl}$ presents basic signal, $\mathbf{XH}_{a,tl}$ indicate the weight matrices of the signal, and μ represents the sparsity constant. Initially, the basis and weight matrices are assigned by random values. The basis matrices $\mathbf{XMRIW}_{a,tl}$ can be created by decreasing the distance between $\mathbf{XMRI}_{a,tl}^{Train}$ and $\mathbf{XMRIW}_{a,tl}\mathbf{XH}_{a,tl} + \mu|\mathbf{XH}_{a,tl}|_1$ using Equation (2) with the assistance of Equation (3) and Equation (4). Likewise, for the source signal $\mathbf{y}(t)$, the basis matrix $\mathbf{YMRIW}_{a,tl}$ is created and concatenated with $\mathbf{XMRIW}_{a,tl}$ like as $\mathbf{XYMRIW}_{a,tl} = [\mathbf{XMRIW}_{a,tl}\ \mathbf{YMRIW}_{a,tl}]$.

3.4. Testing Phase

In the testing phase, the mixed speech signal $\mathbf{z}(t)$ is decayed by using DDDTCWT and takes only approximation coefficients $\mathbf{z}_{a,tl}$, while the detail coefficients are replaced with zero. After applying the STFT to each approximation coefficient sub-band of a mixed signal, the complex spectrum $\mathbf{Z}_{a,tl}(\tau, f)$ is obtained and preserves sign values of RI parts. The magnitude spectrum and the absolute value of RI components are concatenated to generate $\mathbf{ZMRI}_{a,tl}^{Test}$ which are decomposed by using SNMF. The goal is to use SNMF to decompose these spectrums into basis and weight matrices as follows in Equation (8).

$$\mathbf{ZMRI}_{a,tl}^{Test} \approx \mathbf{XYMRIW}_{a,tl}\mathbf{XYH}_{b,tl} + \mu|\mathbf{XYH}_{a,tl}|_1 = [\mathbf{XMRIW}_{a,tl}\ \mathbf{YMRIW}_{a,tl}]\begin{bmatrix}\mathbf{XH}_{a,tl} \\ \mathbf{YH}_{a,tl}\end{bmatrix} + \mu|[\mathbf{XH}_{a,tl}\mathbf{YH}_{a,tl}]|_1 \tag{8}$$

Where $\mathbf{XYH}_{b,tl}$, $\mathbf{XH}_{b,tl}$ and $\mathbf{YH}_{b,tl}$ specify the mixed signal's weight matrices, source signal $\mathbf{x}(t)$, and source signal $\mathbf{y}(t)$, separately. The weight matrix $\mathbf{XYH}_{a,tl}$ can be obtained via SNMF by reducing the distance between $\mathbf{ZMRI}_{a,tl}^{Test}$ and $\mathbf{XYMRIW}_{a,tl}\,\mathbf{XYH}_{a,tl} + \mu\left|\mathbf{XYH}_{a,tl}\right|_1$ using Equation (2) with the assistance of Equation (4), where the preliminary value of $\mathbf{XYH}_{a,tl}$ are assigned by the arbitrary positive numbers, and the values of $\mathbf{XYMRIW}_{b,tl}$ are fixed from the training stage. The first predictable MRI components$\mathbf{XM}_{a,tl}$, $\mathbf{XR}_{a,tl}$, and $\mathbf{XI}_{a,tl}$, respectively, obtained by the following Equations (9-11) for one source signal and similarly obtain $\mathbf{YM}_{a,tl}$, $\mathbf{YR}_{a,tl}$, and $\mathbf{YI}_{a,tl}$ for another source signal.

$$\mathbf{XM}_{a,tl} = \mathbf{XMW}_{a,tl}\mathbf{XH}_{a,tl} \tag{9}$$

$$\mathbf{XR}_{a,tl} = \mathbf{XRW}_{a,tl}\mathbf{XH}_{a,tl} \tag{10}$$

$$\mathbf{XI}_{a,tl} = \mathbf{XIW}_{a,tl}\mathbf{XH}_{a,tl} \tag{11}$$

The magnitude spectrum $\mathbf{ZM}_{\mathbf{b,tl}}$ is not equal to the sum of the preliminary approximation $\mathbf{XM}_{\mathbf{a,tl}}$ and $\mathbf{YM}_{\mathbf{a,tl}}$. To avoid errors, we compute the SBRM using Equation (12) and Equation (13).

$$\overline{\mathbf{XM}}_{\mathbf{a,tl}} = \frac{(\mathbf{XM}_{a,tl})^2}{(\mathbf{XM}_{a,tl})^2+(\mathbf{YM}_{a,tl})^2} \otimes \mathbf{ZM}_{\mathbf{a,tl}} \tag{12}$$

$$\overline{\mathbf{YM}}_{\mathbf{a,tl}} = \frac{(\mathbf{YM}_{a,tl})^2}{(\mathbf{XM}_{a,tl})^2+(\mathbf{YM}_{a,tl})^2} \otimes \mathbf{ZM}_{\mathbf{a,tl}} \tag{13}$$

We also calculate SBRM for RI components. After using the previously saved sign, we multiply it by RI signal estimates. Now, we calculate the phase spectrum. $\mathbf{XP}_{\mathbf{a,tl}}$ from the RI component of one source signal and similarly calculate $\mathbf{YP}_{\mathbf{a,tl}}$. Now, we recombine the estimated phase spectrum $\mathbf{XP}_{\mathbf{a,tl}}$ and $\mathbf{YP}_{\mathbf{a,tl}}$ with an estimated magnitude spectrum $\overline{\mathbf{XM}}_{\mathbf{a,tl}}$ and $\overline{\mathbf{YM}}_{\mathbf{a,tl}}$ to get the modified complex speeches spectrum $\widetilde{\mathbf{X}}_{\mathbf{a,tl}}(\boldsymbol{\tau}, \mathbf{f})$ and $\widetilde{\mathbf{Y}}_{\mathbf{a,tl}}(\boldsymbol{\tau}, \mathbf{f})$ by Equation (14) and Equation (15).

$$\widetilde{\mathbf{X}}_{\mathbf{a,tl}}(\boldsymbol{\tau}, \mathbf{f}) = \overline{\mathbf{XM}}_{\mathbf{a,tl}}\, \mathbf{e}^{\mathbf{iXP}_{a,tl}} \tag{14}$$

$$\widetilde{\mathbf{Y}}_{\mathbf{a,tl}}(\boldsymbol{\tau}, \mathbf{f}) = \overline{\mathbf{YM}}_{\mathbf{a,tl}}\, \mathbf{e}^{\mathbf{iYP}_{a,tl}} \tag{15}$$

The ISTFT is used to convert the altered complex source signals spectrum $\widetilde{\mathbf{X}}_{a,tl}(\tau, f)$ and $\widetilde{\mathbf{Y}}_{a,tl}(\tau, f)$ to the modified sub-band signals $\tilde{\mathbf{x}}_{a,tl}$ and $\tilde{\mathbf{y}}_{a,tl}$. Finally, by applying the IDDDTCWT to the sub-band signals $\tilde{\mathbf{x}}_{a,tl}$ and $\tilde{\mathbf{y}}_{a,tl}$, the expected source speech signals $\tilde{\mathbf{x}}(t)$ and $\tilde{\mathbf{y}}(t)$ are obtained. Algorithms 1 and Algorithm 2 depict the proposed training and testing phases of this system.

Algorithm 1. Training stages for the proposed algorithm

Input: Decomposition level (dl), the number of iterations (k), tree-level (tl), and training sets $\mathbf{x}(\mathbf{t})$ and $\mathbf{y}(\mathbf{t})$.
Output: $\mathbf{XYW}_{p,q}$.

Step 1 : Set p=1, q=1 to tl
Step 2 : Compute the wavelet coefficients using DDDTCWT and take only approximation coefficients. $\mathbf{x}_{p,q} = \boldsymbol{DDDTCWT}(\mathbf{x}(\mathbf{t}))$ and $\mathbf{y}_{p,q} = \boldsymbol{DDDTCWT}(\mathbf{y}(\mathbf{t}))$.
Step 3 : Obtain a complex spectrum by applying STFT. $\mathbf{X}_{p,q} = \boldsymbol{STFT}(\mathbf{x}_{p,q})$ and $\mathbf{Y}_{p,q} = \boldsymbol{STFT}(\mathbf{y}_{p,q})$.
Step 4 : Concatenate the magnitude spectrum and absolute value of the RI component to generate $\mathbf{XMRI}_{p,q}$ and $\mathbf{YMRI}_{p,q}$.
Step 5 : Determine the basis matrices.
For: k=1 to number of iteration
$\mathbf{XW}_{p,q}^{k+1} = \boldsymbol{SNMF}(\mathbf{XMRI}_{p,q}, \mathbf{XH}_{p,q})$ and $\mathbf{YW}_{p,q}^{k+1} = \boldsymbol{SNMF}(\mathbf{YMRI}_{p,q}, \mathbf{YH}_{p,q})$.
End for.
Step 6 : Combine these basis matrices and make $\mathbf{XYW}_{p,q} = [\mathbf{XW}_{p,q}\ \mathbf{YW}_{p,q}]$.
Step 7 j=j+1, go to step 1.

Algorithm 2. Testing stages for the proposed algorithm

Input: Decomposition level (dl), the number of iterations (k), tree-level (tl), mixed-signal ($\mathbf{z}(\mathbf{t})$), combine basis matrices ($\mathbf{XYW}_{p,q}$) learned from the training phase,
Output: Separation of estimated signals $\tilde{\mathbf{x}}(\mathbf{t})$ and $\tilde{\mathbf{y}}(\mathbf{t})$.

Step 1 : Set p=1, q=1 to tl,
Step 2 : Decomposed DDDTCWT and take approximation coefficients, $\mathbf{z}_{p,q} = \boldsymbol{DDDTCWT}(\mathbf{z}(\mathbf{t}))$.
Step 3 : Obtain a complex spectrum, $\mathbf{Z}_{p,q} = \boldsymbol{STFT}(\mathbf{z}_{p,q})$.
Step 4 : Compute magnitude, phase, absolute value, and the sign of RI components from $\mathbf{Z}_{p,q}$ using and concatenating them to prepare $\mathbf{ZMRI}_{p,q}$.
Step 5 : Obtain the weight matrices according to (4).
For: k=1 to test iteration.
$\mathrm{H}\mathbf{Z}_{p,q}^{k+1} = \boldsymbol{SNMF}(\mathbf{ZMRI}_{p,q}, \mathbf{XYW}_{p,q})$.
End for
Step 6 : Estimate the initial magnitude and RI components by using Equation (7-9).
Step 7 : Calculate the sub-band binary ratio masks magnitude, RI components.

Step 8 : Multiply the sign with RI components and obtain the phase spectrum $\mathbf{XP}_{p,q}$ and $\mathbf{YP}_{p,q}$
Step 9 : Apply the phase spectrum $\mathbf{XP}_{p,q}$ and $\mathbf{YP}_{p,q}$ with the estimated source signals magnitude spectrum to obtain the complex spectrum $\tilde{\mathbf{X}}_{p,q}(\boldsymbol{\tau}, \mathbf{f})$ and $\tilde{\mathbf{Y}}_{p,q}(\boldsymbol{\tau}, \mathbf{f})$
Step 10 : Compute the modified sub-band signals.
$\tilde{\mathbf{x}}_{p,q} = \boldsymbol{ISTFT}\left(\tilde{\mathbf{X}}_{p,q}(\mathbf{t}, \mathbf{f})\right)$ and $\tilde{\mathbf{y}}_{p,q} = \boldsymbol{ISTFT}\left(\tilde{\mathbf{y}}_{p,q}(\mathbf{t}, \mathbf{f})\right)$.
Step 11 : q=q+1, go to step 1.
Step 12 : Get predictable source signals $\tilde{\mathbf{x}}(\mathbf{t})$ and $\tilde{\mathbf{y}}(\mathbf{t})$ using IDDDTCWT.

4. Evaluation and Results

4.1. Dataset Description

In this simulation, we extracted speech signals as training and test data from GRID Audio-Visual Corpus [26]. A total of 1000 utterances were made by 34 speakers (18 male and 16 female). For each speaker, we investigate all the sentences. We have used two types of speech signal splitting in this simulation: the first is in the same-gender SS (two cases have male-male and female-female SS), and the second is the opposite-gender SS (one case has male-female SS). In addition, we considered the utterances of eight same-gender speakers to be an experimental group, while the other group had the same utterances of every eight same-gender speakers. We chose thirty-two speakers from the database, including sixteen males and sixteen females. For each speaker, 80% of the sentences in the database were used for training, while the remaining sentences were evaluated for testing. By utilizing a sample rate of 8000 Hz and employing a 512-point STFT with a 50% overlap, the speech signal is transformed into a time-frequency domain. Table 1 represents the specific parameters for the experimental setting.

Table 1. Specific parameters for the experimental setting.

Parameters	Value
Sampling Frequency	8000 Hz
STFT Length	512
Basis Length	70
Wavelet Level	2
Mother Wavelet	dtf2
Training Iteration	70
Testing Iteration	30
Sparsity Level	0.15

4.2. Performance Metrics

The SIR [27], SDR [27], and SAR [27] evaluate the performances of the separated speeaches. The SDR value provides an approximate assessment of the speech quality. The signal is divided according to its input strength, which is determined by the disparity between the input and the replicated signal. Nevertheless, recuperation efficacy is controlled by elevated high SDR scores. Additionally to the SDR, the SIR indicates errors affected by the failure to eliminate the interfering signal throughout the separation process. A higher value of the SIR is associated with an enhanced separation value. The signal-to-artifact ratio (SAR) is a measure of predictable signal quality. The SDR, SIR, and SAR are calculated using Equation (16-18).

i. Source to Distortion Ratio

$$\mathrm{SDR} = 10\log_{10}\frac{\|\mathrm{x}_{\mathrm{target}}\|^2}{\|\mathrm{e}_{\mathrm{interf}}+\mathrm{e}_{\mathrm{noise}}+\mathrm{e}_{\mathrm{artif}}\|^2} \tag{16}$$

ii. Source to Interference Ratio

$$\mathrm{SIR} = 10\log_{10}\frac{\|\mathrm{x}_{\mathrm{target}}\|^2}{\|\mathrm{e}_{\mathrm{interf}}\|^2} \tag{17}$$

iii. Source to Artifact Ratio

$$\mathrm{SAR} = 10\log_{10}\frac{\left\|\mathrm{x}_{\mathrm{target}+\mathrm{e}_{\mathrm{interf}}+\mathrm{e}_{\mathrm{noise}}}\right\|^2}{\|\mathrm{e}_{\mathrm{artif}}\|^2} \tag{18}$$

where $\mathrm{x}_{\mathrm{target}}$, $\mathrm{e}_{\mathrm{interf}}$, $\mathrm{e}_{\mathrm{noise}}$, and, $\mathrm{e}_{\mathrm{artif}}$ represent the targeted source, interference error, perturbation noise, and artifacts error.

4.3. Overall performance of the proposed method

First, the comparison of the proposed algorithm with different models in terms of SDR is shown in Fig. 3. Based on this bar chart, it seems that the offered method outperforms some other current techniques in all cases. For example, when compared to the DTCWT-Noisy_phase method, our approach improves the SDR score of 8.75% for the M1 signal, 11.64% for the M2 signal, 10.05% for the F1 signal, 14.70% for F2 signal, and 9.32% for M signal, 8.88% for the F signal due to different SS scenarios. The majority of speech separation algorithms concentrate on the speech signal's STFT, which only inspects the magnitude spectrum and ignores the phase spectrum. This study takes into account the MRI components and uses DDDTCWT and STFT consecutively with SNMF. Hence, both magnitude, real, and imaginary components were utilized in our proposed method, and the separation performance was improved.

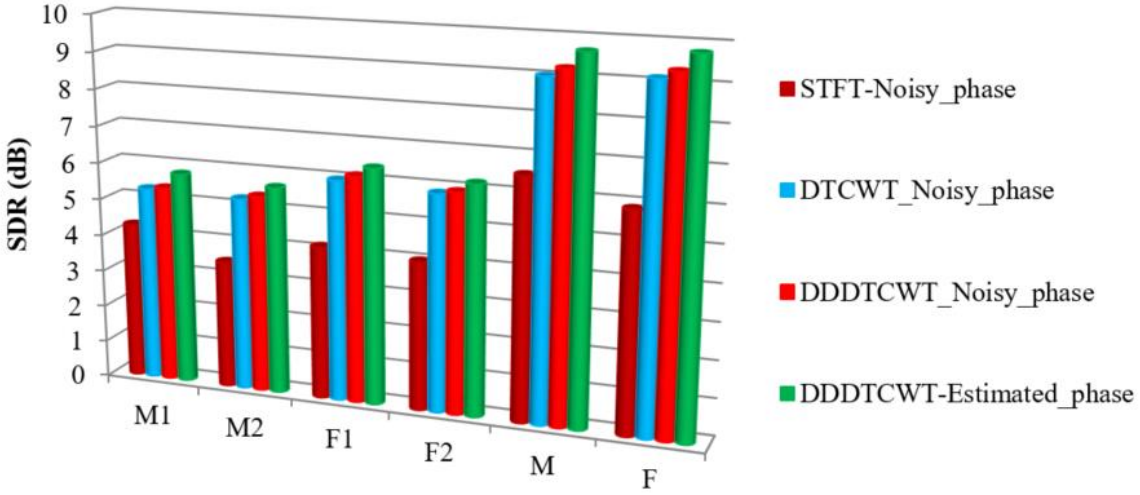

Figure 3. SDR scores of the eight approaches are assessed for cases of the same and opposite gender

Second, we show the examination of the recommended model with the current models regarding SIR in Fig. 4. The figure shows that the proposed algorithm is enhanced for all cases than the other existing methods, specifically STFT-Noisy_phase, and DTCWT-Noisy_phase. The proposed model's SIR values are more advanced for all separation cases than the current models. SIR is improved from 9.32 dB to 9.93 dB for the M1 signal, 9.16 dB to 9.97 dB for the M2 signal, 10.01 dB to 10.64 dB for F1 signal, 9.87 dB to 10.73 dB for F2 signal, 14.57 dB to 15.82 dB for M signal and 13.35 dB to 14.73 dB for F signal using the offered models over DTCWT-Noisy_phase. The STFT typically separates a time domain input signal into discrete frames that are individually considered as stationary. However, we can't distinguish what frequency occurs at what moment of existence, the fragment may not be more stationary. Our proposed technique divides the input signal into small pieces using DDDTCWT, which isolates low and high-frequency components that seem to be more stationary. Separation performance is also improved for the shift-invariant, better direction selectivity, and scheme freedom properties of DDDTCWT.

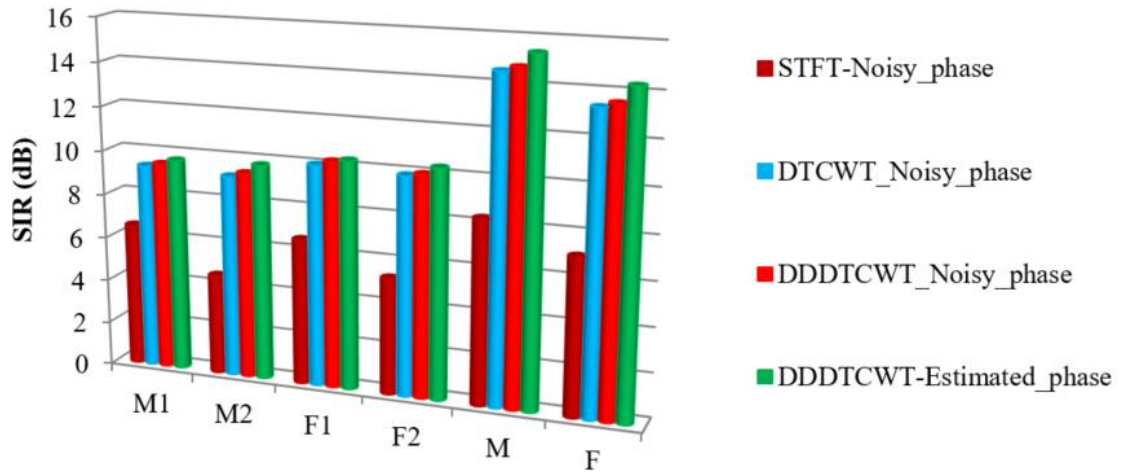

Figure 4. Performance comparison of existing models with the proposed SIR model for same and opposite-gender cases

Third, Fig. 5 depicts a comparative performance analysis of SAR using the recommended method and other existing approaches. We have observed that our recommended DDDTCWT-Estimated_phase can enhance the SAR by 5.70% [= (8.34-7.89)/7.89] for the M1 signal, 7.44% [= (8.52-7.93)/7.93] for the M2 signal, 8.79% [= (8.66-7.96)/7.96] for the F1 signal, 9.60% [= (8.67-7. 91)/7. 91] for the F2 signal, 11.22% [= (10.61-9.54)/9.54] for the M signal, and 13.63% [= (10.75-9.46)/9.46] for the F signal compared to DTCWT-Noisy_phase technique in both same and opposite gender cases. The speech SAR values are improved in every case, implying that the DDDTCWT-Estimated_phase takes care of the speech signal distortion matter after the SS processing.

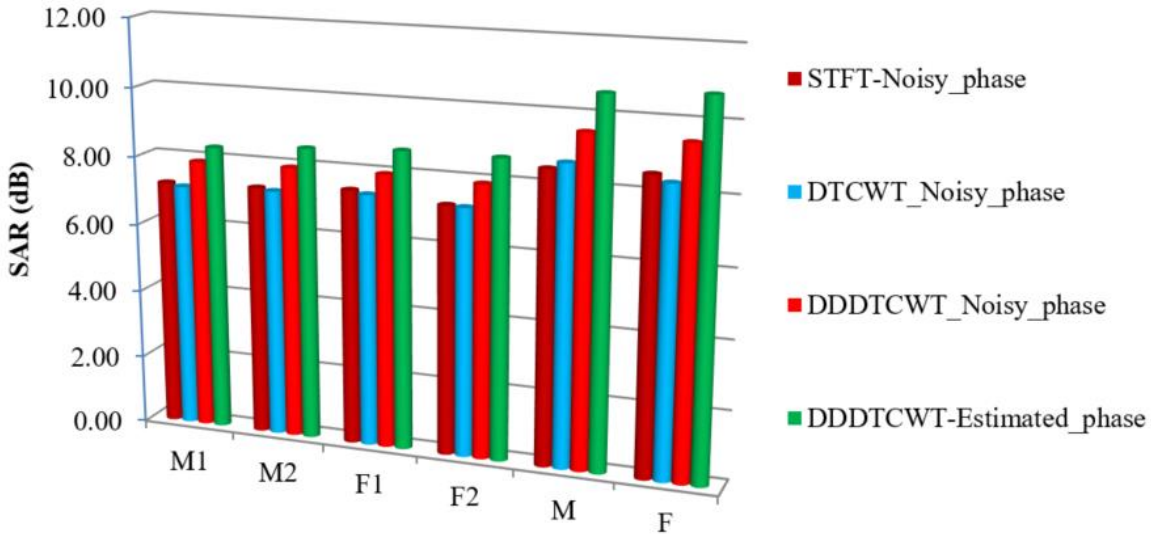

Figure 5. Evaluation of the relative performance of existing models and the proposed SAR model for three different instances

Fig. 6 shows the time-domain waveforms of the ten speech signals, the top two of which are clean speech signals. The rest are estimated speech signals via the STFT-Noisy-phase method, DTCWT-Noisy-phase method, DDDTCWT-Noisy-phase method, and DDDTCWT-Estimated-phase method, respectively. Due to applying the noisy phase introducing some undesired constituents to the predictable speech signals, the SS excellence of existing methods is degraded, as shown in Fig. 6, from the graphs, we see that the proposed method recuperates male and female speech.

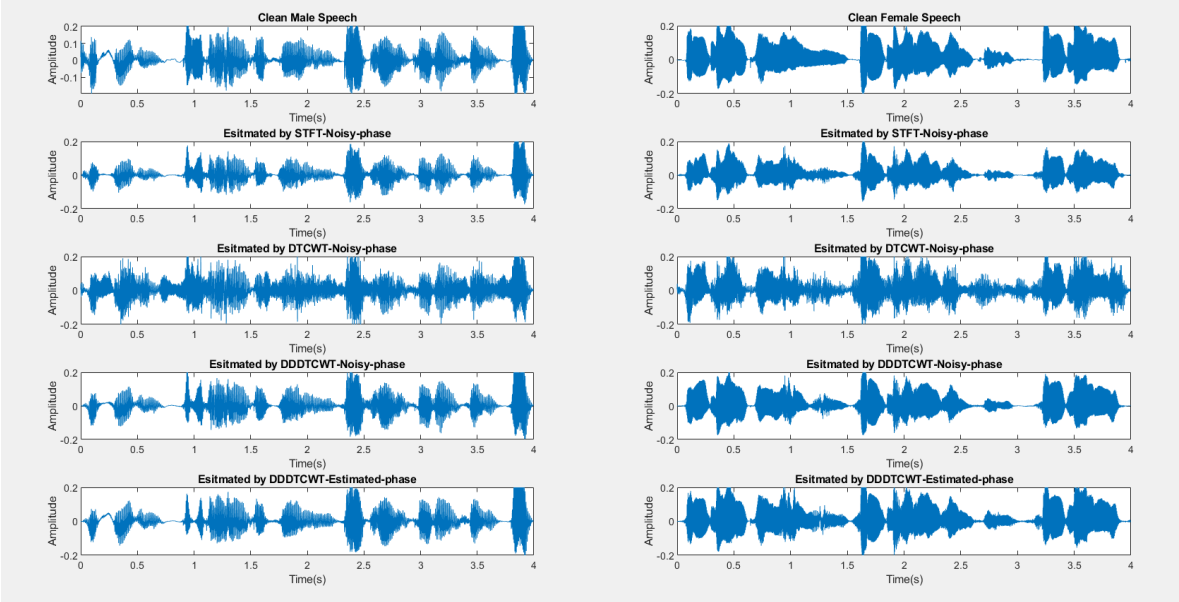

Figure 6. A time-domain waveform of speech, with the x-axis indicating time in seconds and the y-axis indicating amplitude in decibels

Finally, we used the TIMIT database[1] for a mixed speech separation experiment to further confirm the supremacy of advances. The TIMIT database was used to pick 24 speakers (12 male and 12 female) for our study. Each speaker, totalling 240 sentences, says ten sentences. The first eight sentences of ten separate speakers are chosen for training, while the remaining two are selected for testing. We use SDR, SAR, and SIR scores to evaluate the performance of our recommended techniques. The suggested scheme performs superior to the supplementary five techniques that rely on the SDR, SAR, and SIR at opposite gender separation, as shown in Fig. 7.

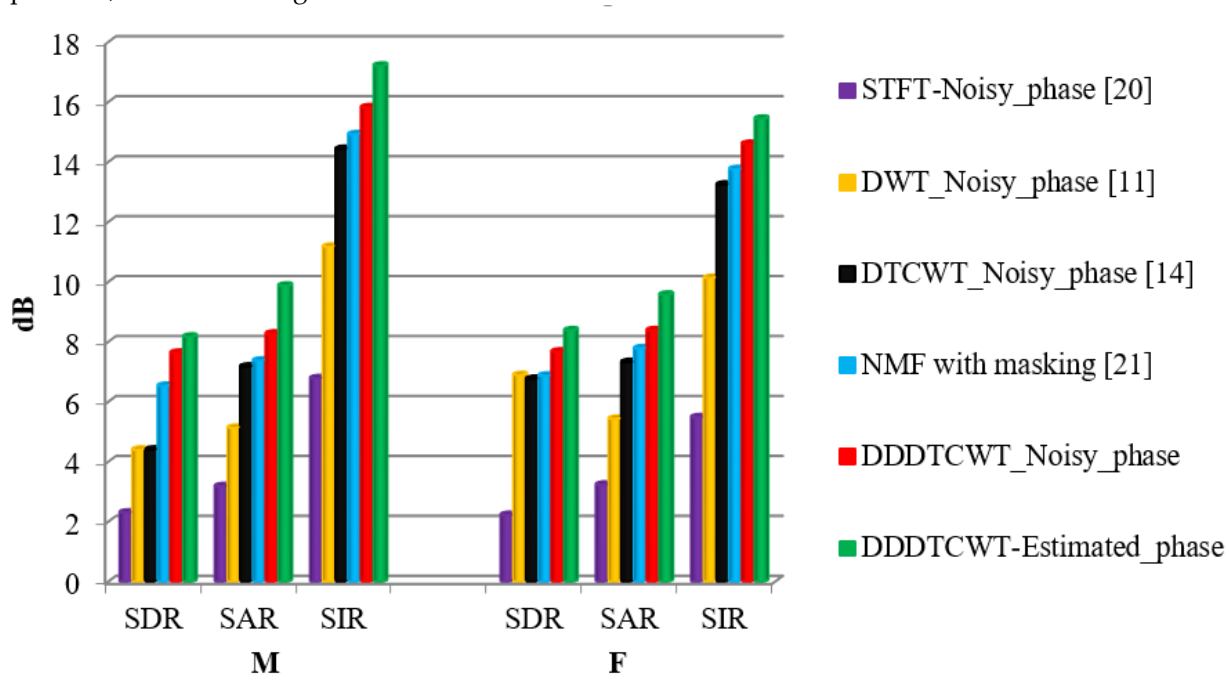

Figure 7. Relative performance assessment of the four models of SDR, SAR, and SIR for the opposite gender case

4. Conclusion

This study proposed a novel technique for separating speech based on double-density dual-transform with join-learning MRI signal portions. In contrast to the usual learning strategy, which only considers the magnitude component, the main goal is to learn the basic vectors while considering the MRI parts together. The DDDTCWT breaks down a time-domain speech signal into high- and low-frequency sub-band sounds. The high-frequency sub-band signal is replaced with zero, and only the low-frequency signal is used. Since high-frequency signal components contain less signal energy, getting rid of high-frequency parts of the input data makes it smaller and faster to use matrix factorization. Consequently, phase information can be considered when utilizing complex domain training targets.

Most researchers merely increase the magnitude spectra, ignoring the phase spectra, and estimating separated speech with noisy phases. We are dealing with the signal's magnitude in addition to its real and imaginary components and are attempting to calculate the phase of the real and imaginary components. The experimental findings utilizing several assessment measures demonstrate that the presented algorithm significantly outperformed the earlier methods from this perspective. We intend to investigate alternative deep neural network-based training and testing algorithms in the future.

References

[1] Po-Sen Huang, Minje Kim, Mark Hasegawa Johnson and Paris Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Print ISSN: 2329-9290, Online ISSN: 2329-9304, pp. 2136-2147, Vol. 23, No. 12, 13 August 2015, Published by IEEE, DOI: 10.1109/TASLP.2015.2468583, Available: https://ieeexplore.ieee.org/document/7194774.

[2] Bo Wu, Kehuang Li, Minglei Yang and Chin-Hui Lee, "A reverberation time aware approach to speech dereverberation based on deep neural networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*,

[1] https://cir.nii.ac.jp/all?q=Linguistic%20Data%20Consortium,%201993.

Print ISSN: 2329-9290, Online ISSN: 2329-9304, pp. 102-111, Vol. 25, No. 1, 31 October 2017, Published by IEEE, DOI: 10.1109/TASLP.2016.2623559, Available: https://ieeexplore.ieee.org/document/7726012.

[3] Rizwan Ullah, Md Shohidul Islam, Md. Imran Hossain, Fazal E. Wahab and Zhongfu Ye, "Single channel speech deriverberation and separation using RPCA and SNMF", *Applied Acoustics*, ISSN: 0003-682X, pp. 107406, Vol. 167, 1 October 2020, Published by Elsevier, DOI: 10.1016/j.apacoust.2020.107406, Available: https://www.sciencedirect.com/science/article/pii/S0003682X20305107.

[4] Kunpeng Wang, Hao Zhou, Jingxiang Cai, Wenna Li and Juan Yao, "Time-domain adaptive attention network for single-channel speech separation", *EURASIP Journal on Audio, Speech, and Music Processing*, Online ISSN: 1687-4722, pp. 1-15, Vol. 2023, No. 1, 11 May 2023, Published by Springer, DOI: 10.1186/s13636-023-00283-w, Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-023-00283-w.

[5] Xiaoming Zhao, Qiang Tuo, Ruosi Guo and Tengteng Kong, "Research on Music Signal Processing Based on a Blind Source Separation Algorithm", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 24-30, Vol. 6, No. 4, 1st October 2022, DOI:10.33166/AETiC.2022.04.003, Available: http://aetic.theiaer.org/archive/v6/v6n4/p3.html.

[6] DeLiang Wang and Guy J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, 1st ed. New York, USA: Wiley-IEEE press, 1 September 2006, Print ISBN: 9780471741091, Online ISBN: 9780470043387, Available: https://ieeexplore.ieee.org/book/5769523.

[7] Aarthi M. Reddy and Bhiksha Raj, "Soft Mask Methods for single channel speaker separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Print ISSN: 2329-9290, Online ISSN: 2329-9304, pp. 1766 - 1776, Vol. 15, No. 6, 23 July 2007, Published by IEEE, DOI: 10.1109/TASL.2007.901310, Available: https://ieeexplore.ieee.org/document/4276763.

[8] Tuomas Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space", In *Proceedings of the INTERSPEECH 2006: Conference of the International Speech Communication Association Interspeech*, Pennsylvania, USA, 17-21 September 2006, DOI: 10.21437/Interspeech.2006-23, Available: https://www.isca-speech.org/archive/pdfs/interspeech_2006/virtanen06_interspeech.pdf.

[9] François G. Germain and Gautham J. Mysore, "Stopping Criteria for Non-Negative Matrix Factorization Based Supervised and Semi-supervised Source Separation", *IEEE Signal Processing Letters*, Print ISSN: 2329-9290, Online ISSN: 1070-9908, pp. 1558-2361, Vol. 21, No. 10, 9 June 2014, Published by IEEE, DOI: 10.1109/LSP.2014.2331981, Available: https://ieeexplore.ieee.org/document/6840338.

[10] Xu LI, Ming TU, Xiaofei WANG, Chao WU, Qiang FU *et al.*, "Single-Channel Speech Separation Based on Non-negative Matrix Factorization and Factorial Conditional Random Field", *Chinese Journal of Electronics*, Print ISSN 1022-4653, Online ISSN 2075-5597, pp. 1063-1070, Vol. 27, No. 5, September 2018, published by IET, DOI: 10.1049/cje.2018.06.016, Available: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cje.2018.06.016.

[11] Yash V. Varshney, Zia A. Abbasi, Musiur R. Abidi and Omar Farooq, "Frequency selection based separation of speech signals with reduced computational time using sparse NMF", *Archives of Acoustics*, Print ISSN: 0137-5075, Online ISSN: 2300-262X, pp. 287-295, Vol. 42, No. 2, 2 November 2017, Published by published by Polish Academy of Sciences, Committee on Acoustics, DOI: 10.1515/aoa-2017-0031, Available: https://acoustics.ippt.pan.pl/index.php/aa/article/view/1878.

[12] Tarek H. Islam, Al Mahmud, Wasim U. Khan and Zhongfu 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*, ISSN: 2079-9292, pp. 353, Vol. 8, No. 3, 22 March 2019, Published by Multidisciplinary Digital Publishing Institute (MDPI), DOI: 10.3390/electronics8030353, Available: https://www.mdpi.com/2079-9292/8/3/353.

[13] Md Shohidul Islam, Yuanyuan Zhu, Md Imran Hossain, Rizwan Ullah and Zhongfu Ye, "Supervised single channel dual domains speech enhancement using sparse non-negative matrix factorization", *Digital Signal Processing*, Print ISSN: 1051-2004, Online ISSN: 1095-4333, pp. 102697, Vol. 100, May 2020, Published by Elsevier, DOI: 10.1016/j.dsp.2020.102697, Available: https://www.sciencedirect.com/science/article/abs/pii/S1051200420300427.

[14] Md Imran Hossain, Md Shohidul Islam, Mst Titasa Khatun, Rizwan Ullah, Asim Masood *et al.*, "Dual-transform source separation using sparse nonnegative matrix factorization", *Circuits, Systems, and Signal Processing*, Print ISSN: 0278-081X, Online ISSN: 1531-5878, pp. 1868-1891, Vol. 40, 23 October 2020, Published by Springer, DOI: 10.1007/s00034-020-01564-x, Available: https://link.springer.com/article/10.1007/s00034-020-01564-x.

[15] Guangzhao Bao, Yangfei Xu and Zhongfu Ye, "Learning a discriminative dictionary for single-channel speech separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Print ISSN: 2329-9290, Online ISSN: 2329-9304, pp. 1130-1138, Vol. 22, No. 7, 29 April 2014, Published by IEEE, DOI: 10.1109/TASLP.2014.2320575, Available: https://ieeexplore.ieee.org/document/6807696.

[16] Yangfei Xu, Guangzhao Bao, Xu Xu and Zhongfu Ye, "Single-channel speech separation using sequential discriminative dictionary learning", *Signal Processing*, Print ISSN: 2329-9290, Online ISSN: 2329-9304, pp. 134-140, Vol. 106, 2 August 2014, Published by IEEE, DOI: 10.1016/j.sigpro.2014.07.012, Available: https://www.sciencedirect.com/science/article/abs/pii/S0165168414003454.

[17] Nitin Kandpal and B. Madhusudan Rao, "Implementation of PCA & ICA for voice recognition and separation of speech", in *Proceedings of the International Conference on Advanced Management Science (ICAMS 2010)*, 9-11 July 2010, Chengdu, China, Vol. 3, DOI: 10.1109/ICAMS.2010.5553181, pp. 536-538, Published by IEEE. Available: https://ieeexplore.ieee.org/abstract/document/5553181.

[18] Sangita Bavkar and Shashikant Sahare, "PCA based single channel speech enhancement method for highly noisy environment", In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 22-25 August 2013, Mysore, India, DOI: 10.1109/ICACCI.2013.6637331, pp. 1103-1107, Published by IEEE, Available: https://ieeexplore.ieee.org/abstract/document/6637331.

[19] Mikkel N. Schmidt, Ole Winther and Lars Kai Hanse, "Bayesian non-negative matrix factorization", In *Independent Component Analysis and Signal Separation, Lecture Notes in Computer Science*, Vol. 5441, Online ISBN: 978-3-642-00599-2, Print ISBN: 978-3-642-00598-5, DOI: 10.1007/978-3-642-00599-2_68, Published by Springer, Berlin, Heidelberg, Available: https://link.springer.com/chapter/10.1007/978-3-642-00599-2_68.

[20] Zi Wang and Fei Sha, "Discriminative non-negative matrix factorization for single-channel speech separation", In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 04-09 May 2014, Florence, Italy, DOI: 10.1109/ICASSP.2014.6854302, pp. 3749-3753, Published by IEEE, 2014, Available: https://ieeexplore.ieee.org/abstract/document/6854302.

[21] MK Prasanna Kumar and R. Kumaraswamy R, "A hybrid model for unsupervised single channel speech separation", *Multimedia Tools and Applications*, Print ISSN: 13807501, Electronic ISSN: 15737721, pp. 1-9, 05 July 2023, DOI: 10.1007/s11042-023-16108-z, Available: https://link.springer.com/article/10.1007/s11042-023-16108-z.

[22] Linhui Sun, Ge Zhu and Pingan Li, "Joint constraint algorithm based on deep neural network with dual outputs for single-channel speech separation", *Signal, Image and Video Processing*, Print ISSN: 18631711, Online ISSN: 18631703, pp. 1387-1395, Vol. 14, No. 7, 12 April 2020, Published By Springer London, DOI: 10.1007/s11760-020-01676-6, Available: https://link.springer.com/article/10.1007/s11760-020-01676-6.

[23] Nasir Saleem, Muhammad I. Khattak, Muhammad Y. Ali and Muhammad Shafi, "Deep neural network for supervised single-channel speech enhancement", *Archives of Acoustics*, Print ISSN: 2300262X, Online ISSN: 01375075, pp. 3-12, Vol. 1, No. 1, 2019, DOI: 10.24425/aoa.2019.126347, Available: https://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-82b7a7cc-98cb-424a-84c7-983c6649c707.

[24] A. S. Yasin, O. N. Pavlova and A. N. Pavlov, "Speech signal filtration using double-density dual-tree complex wavelet transform", *Technical Physics Letters*, Print ISSN: 1063-7850, Online ISSN: 1090-6533, pp. 865-867, Vol. 42, 30 November 2016, published by Springer, DOI: 10.1134/S1063785016080290, Available: https://link.springer.com/article/10.1134/S1063785016080290.

[25] Hanwook Chung, Eric Plourde and Benoit Champagne, "Regularized non-negative matrix factorization with Gaussian mixtures and masking model for speech enhancement", *Speech Communication*, Print ISSN: 0167-6393, Online ISSN: 1872-7182, pp. 18-30, Vol. 87, March 2017, published by Elsevier, DOI: 10.1016/j.specom.2016.11.003, Available: https://www.sciencedirect.com/science/article/abs/pii/S0167639315300145.

[26] Martin Cooke, Jon Barker, Stuart Cunningham and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition", *The Journal of the Acoustical Society of America*, Print ISSN: 0001-4966, Online ISSN: 1520-8524, pp. 2421-2424, Vol. 120, No. 5, 01 November 2006, Published by AIP, DOI: 10.1121/1.2229005, Available: https://pubs.aip.org/asa/jasa/article-abstract/120/5/2421/934379/An-audio-visual-corpus-for-speech-perception-and.

[27] Yi Hu and Philipos C. Loizou, "Evaluation of objective quality measures for speech enhancement", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Print ISSN: 2329-9290, Online ISSN: 2329-9304, pp. 229 - 238, Vol. 16, No. 1, 18 December 2007, Published by IEEE, DOI: 10.1109/TASL.2007.911054, Available: https://ieeexplore.ieee.org/document/4389058.

© 2024 by the author(s). Published by Annals of Emerging Technologies in Computing (AETiC), under the terms and conditions of the Creative Commons Attribution (CC BY) license which can be accessed at http://creativecommons.org/licenses/by/4.0.