Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6655125, 13 pages
https://doi.org/10.1155/2021/6655125

Research Article

An Improved Unsupervised Single-Channel Speech Separation Algorithm for Processing Speech Sensor Signals

Dazhi Jiang,[1] Zhihui He,[1] Yingqing Lin,[1] Yifei Chen,[1] and Linyan Xu[2]

[1] *Department of Computer Science, Shantou University, China 515063*
[2] *Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Italy 20156*

Correspondence should be addressed to Dazhi Jiang; dzjiang@stu.edu.cn

Received 29 December 2020; Revised 24 January 2021; Accepted 1 February 2021; Published 27 February 2021

Academic Editor: Xingsi Xue

As network supporting devices and sensors in the Internet of Things are leaping forward, countless real-world data will be generated for human intelligent applications. Speech sensor networks, an important part of the Internet of Things, have numerous application needs. Indeed, the sensor data can further help intelligent applications to provide higher quality services, whereas this data may involve considerable noise data. Accordingly, speech signal processing method should be urgently implemented to acquire low-noise and effective speech data. Blind source separation and enhancement technique refer to one of the representative methods. However, in the unsupervised complex environment, in the only presence of a single-channel signal, many technical challenges are imposed on achieving single-channel and multiperson mixed speech separation. For this reason, this study develops an unsupervised speech separation method CNMF+JADE, i.e., a hybrid method combined with Convolutional Non-Negative Matrix Factorization and Joint Approximative Diagonalization of Eigenmatrix. Moreover, an adaptive wavelet transform-based speech enhancement technique is proposed, capable of adaptively and effectively enhancing the separated speech signal. The proposed method is aimed at yielding a general and efficient speech processing algorithm for the data acquired by speech sensors. As revealed from the experimental results, in the TIMIT speech sources, the proposed method can effectively extract the target speaker from the mixed speech with a tiny training sample. The algorithm is highly general and robust, capable of technically supporting the processing of speech signal acquired by most speech sensors.

1. Introduction

As information technology is advancing and 5G technology is being popularized, Internet of Things (IoT) devices and sensors will be increasingly created, which will undoubtedly change the way human beings live. Moreover, sensor networks are being progressively studied [1–3]. It is predicted that in the next decade, billions of IoT and sensor devices will generate massive data for applications in smart grid, smart home, electronic health, industry 4.0, etc. It is foreseeable that intelligent speech systems will be critical to the mentioned areas. With the rapid growth of data volume, large-scale problems should be urgently solved effectively [4, 5], while more opportunities are brought. Speech sensor networks, an important part of IoT, will have many application needs. However, in real-world scenarios, the data acquired by speech sensors are often disturbed by noise. Thus, low-noise and effective speech data should be urgently obtained.

With the increasing number of speech sensors, reliable speech separation technology is required [6–8]. High reliable speech separation technology is capable of achieving effective speech recognition, so the needs of human hearing can be satisfied. Speech separation originates from blind source separation (BSS) [9]. The core goal of this technology is to separate the source signal from the measured mixed signal. In the blind source analysis task, the target speech should be separated from the mixed speech in a single channel, which is very difficult to achieve. Single-channel speech separation is a hotspot in the current research. Many algorithms are proposed for single-channel speech separation, but from the current research results, this problem is far from being well solved. We believe that the current challenges are mainly manifested as the following aspects.

(1) Strong noise and unknown number of sources still significantly enhance the performance of BBS. Indeed, most of the existing blind source separation algorithms have achieved ideal performance in high SNR (Signal-to-Noise Ratio) environment. In practical applications, the signal we collected may have been polluted by strong noise. Because of this, many reported algorithms in blind source separation are very likely to obtain poor separation performance and even cannot correctly deal with the severely distorted signal in extreme cases. For the mentioned reason, to obtain robust blind source separation algorithm, a more effective method is required to suppress the impact of noise. In addition, a more difficult problem is that the number of sources is unknown. On the whole, the number of sources should be assumed, whereas in practical applications, information on the number of sources is not available, which cannot be ignored [10]. Accordingly, blind estimation of the number of sources from the received mixed signal cannot effectively obtain the ideal BSS performance

(2) The processing complexity of single-channel speech separation is higher than that of multi-input speech separation. In numerous practical applications, the challenge of blind source separation is that only one sensor is available, namely, SCBSS (single-channel and blind source separation) [11–13]. It uses only a single receiver sensor to receive the observed signal and then uses the signal to recover each source signal. Generative adversarial network (GAN) is an excellent representative of deep learning algorithms and is also used in SCBSS due to its advantages in fitting data distribution (e.g., 1D speech signal separation [14–16]). However, the performance of GAN is limited by the unknown number of source signals, complex forms of dialogue, serious noise pollution, and difficulty in obtaining prior information in advance. Such a type of SCBSS is characterized by unknown number of source signals, complex dialogue form, serious noise pollution, and difficulty in acquiring prior information in advance. To solve this type of problem, unsupervised learning method should be developed, whereas automatic analysis should be extremely difficult to realize based on unsupervised learning method (overall, single-channel speech only requires a single signal source, which is easier to achieve and more realistic than multichannel speech)

(3) The solution to solve BBS problem refers to employing supervised learning mechanism. The more representative is the deep learning method. It has been recently found that deep learning [17, 18] has achieved remarkable success in many speech processing fields with its excellent learning performance. The representative technology is DNN-HMM hybrid structure [19, 20], replacing the conventional acoustic modeling based on GMM and HMM. In single-channel speech separation, a method based on DNNs [21, 22] has been proposed to separate the target speaker from the mixed speech. However, all deep learning algorithms use joint a decoding framework, which requires additional computational complexity. Moreover, deep learning algorithm needs considerable training data, which is difficult to extend to small data sets and unsupervised speech separation scenarios

To reduce the above challenges, an unsupervised speech separation method CNMF+JADE is proposed in this study, i.e., a hybrid method combined with Convolutional Non-Negative Matrix Factorization [23, 24] and Joint Approximative Diagonalization of Eigenmatrix [25]. This study is aimed at performing efficient processing for the highly noisy signal data acquired by the speech sensor to achieve better separation performance. CNMF refers to a nonnegative matrix decomposition method proposed for speech signal processing. The method adopts a 2D time-frequency basis instead of the 1D basis vector in the original nonnegative matrix decomposition, while it ensures the decomposition result to be nonnegative matrix decomposition. Thus, it effectively carries the correlation between local frames of speech signals [26]. JADE is recognized as an adaptive batch independent component optimization algorithm based on multivariate fourth-order cumulative matrix, and it is an effective method for blind source separation. It exploits the feature that mutual accumulation is always zero when signals are independent and builds multiple fourth-order accumulation matrices for multivariate data. Lastly, the mentioned cumulant matrices are jointly diagonalized to solve for the final separated signals [27, 28]. For single-channel signal, CNMF+JADE can effectively separate the overlapped speech including the target speaker. Subsequently, CNMF+JADE with adaptive speech enhancement technology is adopted to further improve the speech quality of the target speaker. To solve the problem of SCBSS, the main innovations can be summarized below.

(1) In this study, CNMF and JADE are combined to solve the problem of single-channel speech separation. The algorithm is appropriate in extracting signals of interest from mixed signals. Specific to SNR (Signal-Noise Ratio), STOI (Short-Time Objective Intelligibility), and PESQ (Perceptual Evaluation of Speech Quality), the proposed CNMF+JADE, as compared with several speech separation methods (CNMF, CNMF +ICA), achieves satisfactory results, especially for single-channel mixed speech

(2) Given the scenario that the speech signal will get worse when the speech signal is enhanced after speech separation, an adaptive method is presented here based on wavelet transform to analyze the speech signal after CNMF+JADE separation, as an attempt to realize selective speech enhancement and increase the efficiency of speech enhancement

The rest of the study is organized as follows. In Section 2, some related studies on the study of single-channel speech separation are presented. In Section 3, the proposed algorithm is elucidated. In Section 4, a specific experimental verification of the performance of the proposed algorithm is presented. Lastly, in Section 5, the conclusion and promising future research directions are drawn.

2. Related Work

As IoT technology is developing, intelligent voice system will have increasingly broad application prospects. In addition, single-channel blind speech separation (SCBSS) technology will arouse wide attention. At present, there are three main directions for SCBSS research:

(1) *Subspace Decomposition-Based Approach* [29]. Methods based on subspace decomposition primarily are aimed at identifying new descriptions. The mentioned new descriptions can often effectively extract perceptive meaningful component sources from complex mixtures [30]. Moreover, new descriptions can eliminate intrusions and reduce signal dimensionality, so redundant components can be avoided. The methods based on subspace decomposition are primarily well established in statistical and transformed data. For instance, in the literature [31], the effectiveness of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) methods in solving subspace decomposition problems has been verified. In fact, methods based on algebraic properties are more often used in dealing with subspace decomposition problems, including Non-negative Matrix Factorization (NMF) [32]. NMF is a classical time-frequency distribution method and is often used for single-channel speech separation [33–37]. Ref [38, 39] highlighted NMF as an unsupervised dictionary-based learning method that effectively helps solve various types of signal separation

(2) *Model-Based Approach*. In the first step of the model-based approach, each speaker in the model scene should be identified, and the gain in the blended frames should be determined. In fact, speaker recognition algorithms have been studied by many authors (e.g., Iroquois [40], Closed loop [41], and Adaptive Speaker Identification (SID) [42]). The next step is to choose an appropriate speech representation. The final step comprises the reconstruction of the speech signal frames, in which separated speech is produced. Overall, the reconstruction usually requires the construction of a hybrid estimator module that enables it to find a sufficient number of representative speech frames from the speaker model to rebuild a meaningful speech signal. However, mixture estimators are capable of significantly complicating the algorithm, so it is difficult to apply in real-time systems

(3) *Computational Auditory Scene Analysis- (CASA-) Based Approach*. CASA runs in two main stages, i.e., segmentation and grouping. The former comprises feature extraction, time-frequency analysis, and multitone tracking, while the latter includes the resynthesis of speech signals. To be specific, pitch tracking is an important technique when CASA is being used for SCBSS problem processing. Jin [43] and Tolonen [44] provided several pitch tracking methods that are used extensively. However, as impacted by the periodic nature of the grouping phase, it can only be limited to voiced speech segments. Moreover, the performance achieved by CASA-based methods tends to be affected by multi-pitch estimation for its dependence on pitch

Over the past few years, with the development of deep learning, researchers have suggested that the nonlinear processing and feature learning capabilities of deep models exhibit significant advantages in solving speech separation problems. For this reason, many models using deep learning for speech separation have been proposed (e.g., Deep Neural Network (DNN), deep stacking, Deep Stack Neural Network (DSN) [45], and other efficient deep learning models [46–50]). In addition, numerous deep learning algorithms have been proposed for single-channel speech separation [51–54]. The reason why deep learning is so effective in addressing with speech separation problems is that the speech separation problem is described as a supervised problem in the deep learning model. Thus, deep learning models can train and learn features from speech signals to effectively separate speech signals.

3. Methodology

In the present section, the methods we use for processing speech data are described, and a new algorithm with high generality and robustness is proposed, aiming to provide a general and efficient speech processing algorithm for the data acquired by speech sensors.

3.1. Speech Separation

(1) *CNMF*. Speech signals exhibit local interframe correlation and global interframe correlation. The conversion of local interframe correlation should consider two aspects, i.e., to ensure the continuity between frames of the converted voice channel spectrum, as well as to remove the source speaker features from the local interframe correlation and make it have the target speaker features. However, the conventional nonnegative matrix factorization does not consider the conversion of local frames. CNMF refers to a proposed nonnegative matrix decomposition method for speech signal processing. The method employs a 2D time-frequency basis instead of the 1D basis vector in the original nonnegative matrix decomposition while ensuring the nonnegativity of the decomposition result. Thus, the correlation between the local frames of the speech signal is carried effectively

The CNMF is expressed as follows:

$$Y \approx \sum_{t=0}^{T-1} A(t) \cdot \overset{t\rightarrow}{X}, \tag{1}$$

where $Y \in M \times N$ and $X \in r \times N$ represent the time-frequency atoms and the corresponding time-varying gain coefficients, respectively. $(\cdot)^{i\rightarrow}$ denotes shifting the encoding matrix X by i units to the right in the form of column vectors and set the leftmost i column to 0.

In other words, the decomposition matrix Y is obtained by convolving a series of nonnegative fundamental matrices A and coefficient matrices X. The functions of CNMF are to find a series of fundamental matrices $A(t)$ and coefficient matrices X and then make the convolution result as close as possible to the target matrix Y.

In addition, the divergence $K-L$ acts as the cost function in CNMF:

$$D\left(Y|\widehat{Y}\right) = \sum_{i,j}\left(Y_{ij}\log\left(\frac{Y_{ij}}{\widehat{Y}_{ij}}\right) - Y_{ij} + \widehat{Y}_{ij}\right), \tag{2}$$

where $\widehat{Y}$ denotes the estimation of $\widehat{Y}$, and

$$\widehat{Y}_{ij} = \left(\sum_{t=0}^{T-1} A(t)\cdot \overset{t\rightarrow}{X}\right)_{ij}. \tag{3}$$

$K-L$ makes the maximum log-likelihood solution of solving the nonnegative matrices $A(t)$ and X under the Poisson noise assumption to describe the degree of approximation of $\widehat{Y}$ with respect to Y. The iterative function can be defined as follows.

$$X = X \otimes \frac{A(t)^T \cdot \left[\overset{\leftarrow t}{Y} / \widehat{Y}\right]}{A(t)^T \cdot E}, \tag{4}$$

$$A(t) = A(t) \otimes \frac{\left(Y/\widehat{Y}\right)\cdot \overset{t\rightarrow}{X}^T}{E \cdot \overset{t\rightarrow}{X}^T}, \tag{5}$$

where E indicating the matrix with all elements of 1 and $\otimes$ is the matrix element multiplication operator. When $T = 1$, i.e., $t = T - 1$ is 0, it will degenerate into the basic NMF decomposition. For each t, there is a basic matrix $A(t)$ corresponding to it.

3.1.1. JADE. The Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm is an adaptive batch independent component optimization algorithm based on multivariate fourth-order cumulative matrices and an effective method for blind source separation. JADE mainly uses the diagonalization of Jacobi matrix to find the independent components, as an attempt to achieve the identification and separation of signals. Based on the characteristics of JADE mentioned above, JADE is introduced to effectively separate the acquired speech signals.

JADE algorithm first spheres the observed signal using an $n \times m$ spherization matrix to obtain the observation vector $u = [u_1, u_2, \cdots, u_N]^T$ for N channels. Then, let M be any $N \times N$ matrix, then the definition of the four-dimensional cumulant matrix $Q_u(M)$ of u is:

$$[Q_u(M)]_{ij} = \sum_{k=1}^{N}\sum_{l=1}^{N} K_{ijkl}(u) m_{kl}, i, j = 1, 2, \cdots, N, \tag{6}$$

where $K_{ijkl}(u)$ denotes the fourth-order cumulant of the i, j, k, and l components in the vector.

3.1.2. CNMF+JADE. However, in the same channel spectral matrix Y, the final $A(t)$ and X obtained by CNMF analysis are not the same when the initial values of $A(t)$ and X are different, i.e., the same time-frequency spectral matrix Y has multiple combinations of time-frequency bases and coding matrices. For the mentioned reason, if the parallel channel spectral matrices of the source and target speakers are analyzed independently by convolutional nonnegative matrix decomposition, the same encoding matrix that characterizes the content information is not ensured to be obtained. From the analysis described in Section 3.1.1, JADE is known as an adaptive batch independent component optimization algorithm based on multivariate fourth-order cumulative matrices and an effective method for blind source separation, capable of effectively identifying and separating signal, which achieves the obtained signals as identical as possible.

Accordingly, to efficiently process the speech signals collected by the speech sensors, a single-channel speech separation algorithm combining CNMF and JADE is proposed. The secondary separation process is performed on the speech signal separated by CNMF based on JADE. The role of CNMF+JADE algorithm is to separate the single-channel mixed speech and lastly acquire the separated speech signal of all speakers in the mixed speech. The algorithm exhibits strong generality and robustness, capable of technically supporting the processing of speech signals collected by most speech sensors. For instance, in the literature [55], several applications (e.g., beamforming, automatic camera steering, robotics, and surveillance) are processed with the speech separation method. In [56], a speech signal separation method is adopted for speech separation of noisy robust speech translation for general-purpose smart devices. It is foreseen that speech separation techniques are also critical to future applications of IoT technologies (e.g., driverless, smart home, and other applications involving sound conduction functions). For this reason, it is of great value and significance to propose more efficient speech separation algorithms (e.g., CNMF+JADE) as proposed in this study.

Lastly, the CNMF+JADE algorithm is described as follows.

The proposed algorithm is written in Algorithm 1, where $t_1, t_2, \cdots, t_N$ represent the set of all the pure speech signal data of the speaker waiting to be separated, $o_1, o_2, \cdots o_{N-1}$ denote the set of all the mixed speech employed as the training set, O is a mixed speech waiting to be separated, R_i is a random

```
Input: Speech signal dataset, $t_1, t_2, \cdots, t_N$, $o_1, o_2, \cdots o_{N-1}$ and $O$.
1:  Initialize each parameter and variable:
      $T$=$t_1, t_2, \cdots, t_N$, expresses the set of all the pure speech signal data of the speaker waiting to be separated,
      $H$=$o_1, o_2, \cdots o_{N-1}$, expresses the set of all the mixed speech that is used as the training set,
      $O$ denotes a mixed speech waiting to be separated,
      $R_i$ is a random matrix.
2:  while $i < N$ do
3:      The speech data with the identical subscript $t_i$and$o_io_i$ from the datasets T and H are selected to train CNMF.
4:      The trained CNMF is employed to separate the mixed speech dataset $O$ to determine $\hat{s}_i$ and $\widehat{O}_i$.
5:      The two speech signals acquired from 4 are mixed to obtain a two-channel speech signal and stored in $R_i$.
6:      A secondary separation is conducted by adopting JADE to obtain $\hat{s}_i$ and $\widehat{O}_i$ from $R_i$.
7:      $\widehat{O}_i$ is used as the speech signal to be separated in the next round, and $t_i$ and $o_i$ are removed from the data sets $T$ and $H$.
8:      Obtain the final separated speech signal.
9:      i = i +1.
10: end while.
Output: All of speaker's speech signals $s_1, s_2, \cdots, s_N$.
```

Algorithm 1: CNMF+JADE description.

matrix, and N represents the number of speech signals, i.e., the number of speakers. o_i denotes the corresponding speaker, as expressed in the dataset O.

$$o_i = O - \sum_{k=1}^{i} t_k, i = 1, 2, \cdots, N - 1. \tag{7}$$

In fact, $o_1, o_2, \cdots o_{N-1}$ are very costly and difficult to obtain. Thus, in experiments, a speech signal different from the current target speaker is generally selected randomly from the dataset O to train CNMF. Although the results obtained by this approach are slightly degraded, the proposed algorithm can be applied to more general range.

In addition, R_i, mentioned in Table 1, is a 2×2 matrix, which is represented as follows:

$$S_i = R_i * \left[\hat{s}_i \,; \widehat{O}_i\right], i = 1, 2, \cdots, N, \tag{8}$$

where $\hat{s}_i\widehat{O}_i$ denotes the speech signal of the target speaker and $\widehat{O}_i$ denotes the set of speech signals obtained after the separation of all speakers.

According to the defects of some existing single-channel speech separation methods, a new algorithm combining CNMF and JADE is proposed in this study. The CNMF is first trained using the training speech signal, and the trained CNMF is used to separate the mixed speech. Next, the separated speech signals are mixed, and the secondary separation is conducted by using JADE. In the next section, simulation experiments are performed to verify the performance of the proposed algorithm and compare it with several other algorithms.

3.2. Speech Enhancement. Some noise usually remains in the target speaker's speech after speech separation, and the interference of noise will inevitably reduce the quality and intelligibility of speech. For the mentioned reason, suppressing the background noise and extracting the pure speech becomes an important part of the speech processing process. Speech enhancement techniques should be used to enhance the target signal after speech signal separation. The conventional single-channel speech enhancement techniques comprise checkpoints [57], Wiener filtering [58], Kalman filtering [59], wavelet transform [60], and so on.

However, as reported by some existing studies, wavelet transform has more significant advantages in single-channel speech signal enhancement. Moreover, the experiments in this study prove this point. Wavelet transform is another landmark technique after Fourier transform. Wavelet transform inherits the advantages of Fourier transform while overcoming its defects. It is an ideal tool for signal time-frequency analysis and processing. One of the features of the wavelet transform in signal processing is that the transform can make certain aspects of the signal more prominent, so it is enabled to highlight signal details when processing the signal and thus extract the effective signal.

Accordingly, based on the above motivation, we will use wavelet transform as the speech signal enhancement technique in this study and propose a more effective adaptive wavelet transform to enhance the extracted signal.

In the following, the wavelet transform and the adaptive wavelet transform technique proposed in this study are introduced.

3.2.1. Speech Enhancement Based on Wavelet Transform. In the present section, we introduce the wavelet transform to enhance the sensor speech signal. The principle of wavelet transform is described below.

Set $L^2(R)$ as a square integrable space, and $\phi(t) \in L^2(t)$, if its Fourier transform satisfies Eq. (9) as follows:

$$C_\phi = \int_R \frac{|\phi(\omega)|^2}{|\omega|} d\omega < \infty. \tag{9}$$

$\phi(\omega)$ denotes a basic wavelet or a mother wavelet.

Table 1: Simulate the voice signal data acquired in different scenarios.

Scene	Number	Speaker	Target
2 speakers	a	1 female +1 female	
	b	1 female +1 male	
	c	1 male +1 male	
3 sparkers	d	1 female +2 males	1 female
	e	3 females	1 female
	f	2 females + male	1 female
	g	3 males	1 male

After the mother wavelet $\phi(t)$ is scaled and translated by a real pair (a, b), where $a, b \in R$, $a \neq 0$, a cluster function can be yielded:

$$\phi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\phi\left(\frac{t-b}{a}\right), a, b \in R \,; a \neq 0. \tag{10}$$

This cluster function denotes a wavelet basis function, where a represents the scaling factor and b denotes the translation factor. $\phi((t-b)/a)$ represents a window function whose window size is fixed but its shape can be changed. According to this characteristic, the wavelet transform is characterized by multiresolution analysis. $1/\sqrt{|a|}$ is a normalization factor, so the wavelets are enabled to have the same energy at different scales.

Signal processing based on wavelet domain is one of the main methods of speech signal processing. Wavelet transform has the characteristics of multiresolution, low entropy, and decorrelation, enabling the wavelet transform to show significant advantages in speech signals processing. Moreover, considerable wavelet bases can theoretically handle different scenarios, so the wavelet transform is significantly useful for speech signal processing.

The main process of wavelet transform denoising is shown in Figure 1, which well demonstrates the process.

3.2.2. Speech Enhancement Based on Adaptive Wavelet Transform. As suggested from the results of the experiments of this study, the quality of the enhanced speech signal may be reduced when the speech signal is enhanced after speech separation. This result proves that the speech enhancement algorithm cannot denoise properly on all noisy speech. In the present section, this study presents an adaptive method based on wavelet transform to analyze the CNMF+JADE separated speech signals and try to achieve selective speech enhancement, that is, before speech enhancement, automatic filtering those speech segments may cause quality degradation. As indicated from the analysis of the speech signal after separation and the speech after wavelet transform, under the significant difference between the separated speeches, the quality will reduce while increase with the wavelet transform. Based on the mentioned findings, the following method is developed to process adaptive judgment before speech enhancement.

$$\begin{gathered}
\text{is\_enhance} = \begin{cases} 1, \operatorname{disp}\left(\hat{s}_i, \hat{O}_i\right) \leq p * \left(\operatorname{disp}\left(\hat{s}_i, \hat{O}_{i-1}\right) + \operatorname{disp}\left(\hat{O}_i, \hat{O}_{i-1}\right)\right)/2 \\ 0, \qquad \text{otherwise}, \end{cases} \\
i = 1, 2, \cdots, N, \\
O_{i-1} = O_i + s_i + l, \\
O_0 = O, \\
O_N = s_N,
\end{gathered} \tag{11}$$

where s_i denotes the ith target speaker speech signal after CNMF+JADE separation. O_i is the mixed speech signal after the CNMF+JADE separation on the mixed speech O_{i-1}. $\hat{s}_i$ and $\hat{O}_i$, respectively, express the Gaussian Mixture Model (GMM) [61, 62] of s_i and O_i. l indicates the loss during the separation process. N represents the number of speakers included in the mixed signal. p is the scaling factor, and the value is [1, 1.2].

$\operatorname{disp}(\cdot)$ represents the GMM distance calculation formula, as defined below:

$$\operatorname{disp}(A, B) = \sum_{i=1}^{M} WA_i \left(\sum_{j=1}^{M} WB_j d_{AB}(i, j)\right). \tag{12}$$

The function of $\operatorname{disp}(\cdot)$ is to measure the dispersion between A and B, i.e., the coupling degree, and W is the weight.

Equation (11) can be explained as under the low coupling between and obtained by CNMF+JADE separation, no further speech enhancement is performed. In other words, under the 0 value obtained from Eq. (11), it is considered that the better the separation effect of the CNMF+JADE algorithm, the less noise the separated speech will contain, and then, further speech enhancement may be counterproductive. Furthermore, under the value of 1, the experimental wavelet transform is considered to be required for separation again.

Equation (11) adaptively determines which separated signals should be enhanced again and which ones do not, so the separated speech signals can be effectively optimized.

Finally, Figure 2 illustrates the flow of the whole algorithm.

4. Experiment Verification

As impacted by the limitations of the experimental conditions, in the present section, a sensor will be simulated to acquire speech data in a speech scene. The basic data used in the experiments originate from an acoustic-phonetic continuous speech corpus constructed in collaboration with Texas Instruments, MIT, and SRI International, i.e., the TIMIT dataset. The TIMIT dataset exhibits a speech sampling frequency of 16 kHz and comprises a total of 6300 sentences spoken by 630 individuals from eight major dialect regions in the United States. All sentences were manually segmented at the phoneme level (phone level) and then labeled. 70% of the speakers were male, and the speakers were

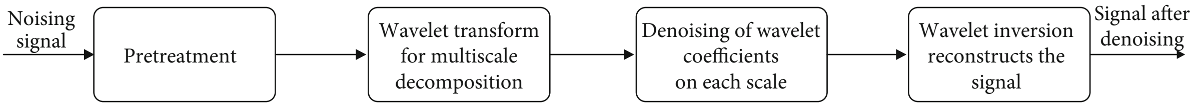
Noising signal
Pretreatment
Wavelet transform for multiscale decomposition
Denoising of wavelet coefficients on each scale
Wavelet inversion reconstructs the signal
Signal after denoising

Figure 1: Wavelet denoising process diagram.

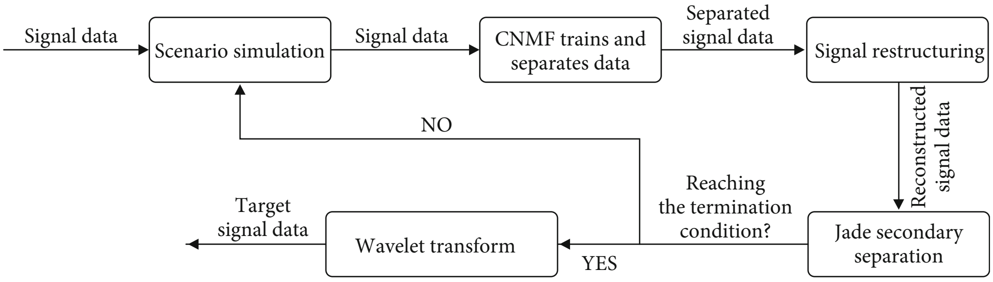

Figure 2: Flow chart of CNMF+JADE+wavelet transform.

primarily white adults. Next, multiple scenarios are simulated, and the speech signals are mixed according to the different scenarios.

For experimental design, in the first part of the experiment, different algorithms are used to separate the speech signals, and then, the signals are analyzed and compared with the algorithm proposed in this study to show that the CNMF +JADE algorithm proposed here can apply to the analysis and processing of the signal data collected by speech sensors. Subsequently, in the second part of the experiments, the performance of several single-channel speech enhancement techniques is verified, and the ability of the adaptive wavelet transform technique proposed in this study to effectively enhance the separated speech signals is experimentally verified, proving the effectiveness of the proposed method. In the following, the experiments are elucidated.

According to Table 1, the case of a speech was simulated, and two scenarios were set up. Scenario I contains two speakers and sets three specific scenarios. Scenario II contains three speakers, one of whom is the target speaker, and sets four specific scenarios. All the scenarios are set up with numbers *a-g*.

In addition, three scientific evaluation metrics are adopted to scientifically evaluate the quality of the separated speech signal. The three evaluation metrics introduced and their descriptions are elucidated below:

(1) Signal-Noise Ratio (SNR) [63] is the ratio between the valid signal and the invalid signal (noise signal). The larger the ratio, the greater the proportion of valid signals will be, and the purer the signal will be

(2) Perceptual Evaluation of Speech Quality (PESQ) [64] is an objective, full-reference speech quality assessment method that considers the subjective perception of human speech signals and can provide a subjective predictive value for objective speech quality assessment, which is recognized as an objective reflection of subjective evaluation. The PESQ score ranges between [-0.5,4.5], and a higher score indicates better speech quality after separation

(3) Short-Time Objective Intelligibility (STOI) [65], like PESQ, refers to a common objective evaluation method that conforms to the human auditory system for speech quality evaluation. It represents the actual intelligibility of speech, with the value ranging between [0,1]. If the value is closer to 1, the more easily the separated speech will be understood, and the higher the intelligibility will be

4.1. Speaker Separation. In the present section, simulation experiments are performed to verify the effectiveness of the proposed algorithm. According to the way of sound mixing, the speech signal separation falls to mono and multichannel speech separation. Since multichannel speech signals involve more available knowledge than monophonic speech signals, multichannel speech signals are simpler to process. The common multichannel speech separation algorithms are mainly based on Independent Components Analysis (ICA) and have shown better performance. For this reason, in the present section of experiments, we selected ICA as the comparison algorithm for speech separation. However, it should be noted that our simulation experiments are based on single-sensor hybrid speech separation, which does not satisfy the application of the ICA algorithm. Thus, in this part of the experiments, we extend the ICA algorithm by combining ICA with CNMF so that it can be applied to but-channel speech separation and compare it with the algorithm proposed in this study. Lastly, the specific methods used in this study are CNMF, CNMF+ICA, and CNMF+JADE.

Table 2 shows the results of the experiments by employing different separation methods, *a-g* corresponding to several dialogue scenarios simulated above in turn. The values in the table represent the evaluated results of the target speaker's speech and the original pure speech with the corresponding methods, in which the data corresponding to the MIX method refer to the data of the three metrics corresponding to the original mixed pure speech, and the later data are the results achieved with the three methods CNMF, CNMF+ICA, and CNMF+JADE, respectively. The best experimental results in each scenario are marked in italics.

Table 2: The results of different speech separation methods (SNR, PESQ, STOI).

					Voice			
Method	*Index*	*a*	*b*	*c*	*d*	*e*	*f*	*g*
	SNR	1.64	5.45	-0.77	1.83	-0.61	0.08	-2.07
MIX	*PESQ*	2.27	2.05	2.43	1.58	1.84	1.75	1.73
	STOI	0.87	0.87	0.74	0.76	0.75	0.78	0.60
	SNR	9.31	7.93	8.46	*8.52*	5.94	6.38	4.89
CNMF	*PESQ*	2.85	2.21	2.49	1.85	1.99	1.97	2.08
	STOI	0.85	0.85	0.77	0.80	0.78	0.78	0.73
	SNR	9.28	10.62	5.42	6.88	*7.19*	5.71	2.63
CNMF + ICA	*PESQ*	2.10	1.84	1.80	1.78	1.80	1.66	1.78
	STOI	0.94	*0.93*	0.89	*0.89*	0.85	0.88	0.62
	SNR	*13.10*	*9.20*	*11.19*	8.44	7.03	*8.32*	*5.68*
CNMF + JADE	*PESQ*	*3.02*	*2.40*	*2.69*	*2.05*	*2.20*	*2.26*	*2.35*
	STOI	*0.95*	0.90	*0.92*	0.88	*0.85*	*0.89*	*0.74*

From the experimental results in the table, we can find that the speech signals processed by all methods are significantly improved compared to the original mixed speech MIX. In addition, the CNMF+JADE algorithm proposed in this study achieves the best experimental results in almost all scenarios; among the 7 scenarios and 21 metrics, only 4 metrics are worse than the experimental results of other methods (CNMF+ICA), which are the SNR and STOI results of scenario *b* and STOI results of scenario *d*. Moreover, it can be seen that the experimental results evaluated using PESQ are all better than those calculated by several other algorithms, which fully demonstrates the effectiveness of the proposed algorithm.

First, the proposed CNMF+JADE algorithm is compared with the CNMF algorithm, and all experimental results are found to outperform those of CNMF, which demonstrates that combining JADE with CNMF is effective. Subsequently, as revealed from the comparison with the CNMF+ICA algorithm, almost all the results are better than those achieved by CNMF+ICA, indicating that combining JADE with CNMF is a purposeful combination and more promising. The combined experimental results fully illustrate the effectiveness of the proposed algorithm.

4.2. The First Experiment Verification for Enhancement. In this part of the experiments, the performance of several conventional single-channel speech signal enhancement techniques is compared. The separated signal complies with the signal of the target speaker obtained from the CNMF +JADE method in Section 4.1. Moreover, the CNMF+JADE method is the method proposed in this study. Subsequently, the target speech signal is enhanced with the four speech enhancement methods separately, and lastly, the enhanced speech signal is evaluated with SNR, PRSQ, and STOI. The experimentally achieved results are listed in Table 3, where the experimental results of the CNMF+JADE method represent the experimental results to be compared. Likewise, *a*-*g* columns correspond to the various scenarios in Table 1, in which each method is evaluated with three evaluation metrics.

First, comparing the four conventional single-channel speech enhancement methods, it can be found that the algorithm using wavelet transform as the speech enhancement method exhibits the optimal performance among the four conventional speech enhancement methods. As suggested by the experimental results achieved with SNR as the evaluation index, the wavelet transform achieves the optimal results in all seven scenarios. For the experimental results achieved with STOI as the evaluation index, six scenes also achieve the optimal results, and only the experimental results of scenario *a* are slightly lower than those of the wiener filtering method, and the differences are slight, 0.82 and 0.83, respectively. Specific to the experimental results achieved with PESQ as the evaluation index, four of the seven scenes achieve the optimal results. As indicated from the comprehensive experimental results, the enhancement of the speech signal obtained by separating CNMF+JADE algorithm using wavelet transform is very effective. For the mentioned reason, this is one of the motivations for choosing wavelet transform as the speech enhancement method in this study.

In addition, the results of the experiments in which the wavelet transform method is used are compared with the results of the experiments in which the speech enhancement method is not used. It can be found that not all the speech quality is enhanced after speech enhancement. For instance, specific to scenario *a*, the speech quality obtained after using the wavelet transform method decreases in all cases. For the experimental results achieved by using wavelet transform as the speech enhancement method, a total of 11 results out of 7 scenes and 21 results are better than the experimentally achieved results without the speech enhancement method.

For this reason, it can be concluded that the purpose of speech enhancement is to remove the noise in the speech segment and thus improve the quality of speech. However, during speech enhancement, the speech signal is corrupted to a certain extent, so the speech quality turns out to be not necessarily better after speech enhancement.

Table 3: The results of the enhancement methods (SNR, PRSQ, STOI).

Method	Index	Scene *a*	*b*	*c*	*d*	*e*	*f*	*g*
CNMF + JADE	*SNR*	13.10	9.20	11.19	8.44	7.03	8.32	5.68
	PESQ	3.02	2.40	2.69	2.05	2.20	2.26	2.35
	STOI	0.95	0.90	0.92	0.88	0.85	0.89	0.74
CNMF + JADE+ spectral subtraction	*SNR*	5.22	5.36	6.16	4.76	4.07	4.76	3.55
	PESQ	1.97	1.95	1.70	1.89	1.88	1.89	1.27
	STOI	0.79	0.82	0.70	0.79	0.72	0.80	0.75
CNMF + JADE+ Wiener filtering	*SNR*	-2.24	-2.04	-1.02	-1.96	-2.13	-1.96	-1.62
	PESQ	2.36	2.33	2.96	2.13	2.19	2.13	2.49
	STOI	0.73	0.77	0.72	0.72	0.69	0.72	0.72
CNMF + JADE+ Kalman filtering	*SNR*	3.27	3.42	3.19	3.05	2.84	4.0	2.05
	PESQ	2.03	2.25	2.05	2.00	1.97	2.00	2.04
	STOI	0.83	0.88	0.69	0.82	0.78	0.83	0.57
CNMF + JADE+ wavelet transform	*SNR*	12.02	11.10	15.9	8.35	7.23	8.37	7.40
	PESQ	2.49	2.46	1.96	1.70	2.25	2.15	1.90
	STOI	0.82	0.92	0.81	0.84	0.88	0.91	0.76

Thus, it is very important and necessary to adaptively select the speech signals that should be enhanced, instead of blindly enhancing all signals. For this reason, this study proposes an adaptive wavelet transform method that adaptively selects the enhanced speech signals and filters out the speech signals that are not required to be enhanced. The specific experimental validation is presented in the next section.

4.3. The Second Experiment Verification for Enhancement.

In this part of the experiments, the adaptive wavelet transform enhancement method proposed in this study is validated. Again, the enhanced speech signal is acquired from the speech signal obtained after separation using the CNMF +JADE method. Moreover, the experimental results of the three metrics are verified separately. The achieved experimental results are listed in Tables 4–6, which fall to three parts, i.e., CNMF+JADE for the experimental results without enhancement and CNMF+JADE+wavelet transform for the experimental results with wavelet transform. Lastly, the adaptive wavelet transform method proposed here is adopted to evaluate whether the speech signal should be enhanced in each scene. From the experimental results in Tables 4–6, we can see that 0 is the experimental result without enhancement, and the corresponding experimental results with wavelet transform enhancement have decreased. 1 is the experimental result with enhancement, and the corresponding experimental results with wavelet transform enhancement have improved.

It is demonstrated through experiments that our adaptive judgment method can filter out the speech segments whose quality will be degraded after wavelet transform. As revealed from the results, the adaptive wavelet transform speech enhancement method proposed in this study can automatically filter the speech segments that are not suitable for speech enhancement, thus effectively improving the quality of the final speech signal.

4.4. Compared with the Deep Learning.

In recent years, with the development of deep learning, researchers have noticed that the nonlinear processing and feature learning capabilities of deep models have significant advantages in addressing speech separation problems. Thus, in this part of the experiments, we implemented a cyclic stacking neural network (Ref [66]) to perform separation processing of the acquired speech signals. In Ref, the speech separation results of various deep neural networks are compared, which are close to the work in this study. We use two metrics, PESQ and STOI, to evaluate the quality of the separated speech signal to compare the performance of the proposed algorithm with deep learning algorithms. Comparing the results of the proposed speech separation methods, we can dig out the advantages and disadvantages of the shallow and deep models.

The experimental results of the proposed algorithm and the deep learning algorithm are shown in Table 7. From the experimental results, we can see that there is still a gap between the method proposed in this study and the deep learning method. In terms of PESQ index, the improvement of RDSN is obviously better than the method in this study. As indicated from the experimental results achieved with STOI as the evaluation index, the optimal value of the proposed method in this study is 0.106, which is the same as the experimental result of DDN, and the difference with the experimental result of RDSN is not much, only 0.006.

As indicated from a comprehensive analysis of the experimental results, the deep model outperforms the shallow model in the supervised case. However, the deep model requires considerable training data, and a large amount of speech data are very difficult to obtain. In addition, the deep model is more expensive to train, and it is difficult to achieve

Table 4: The experiment of adaptive judgment speech enhancement (SNR).

	Scene						
Method	***a***	***b***	***c***	***d***	***e***	***f***	***g***
CNMF + JADE	13.10	9.20	11.19	8.44	7.03	8.32	5.68
CNMF + JADE+ wavelet transform	12.02	11.10	15.9	8.35	7.23	8.37	7.4
Adaptive speech enhancement judgment	0	1	1	0	1	1	1

Table 5: The experiment of adaptive judgment speech enhancement (PESQ).

	Scene						
Method	***a***	***b***	***c***	***d***	***e***	***f***	***g***
CNMF + JADE	3.02	2.40	2.69	2.05	2.20	2.26	2.35
CNMF + JADE+ wavelet transform	2.49	2.46	1.96	1.70	2.25	2.15	1.90
Adaptive speech enhancement judgment	0	1	1	0	1	1	1

Table 6: The experiment of adaptive judgment speech enhancement (STOI).

	Scene						
Method	***a***	***b***	***c***	***d***	***e***	***f***	***g***
CNMF + JADE	0.95	0.90	0.92	0.88	0.85	0.89	0.74
CNMF + JADE+ wavelet transform	0.82	0.92	0.81	0.84	0.88	0.91	0.76
Adaptive speech enhancement judgment	0	1	1	0	1	1	1

Table 7: Efficiency comparison of speech separation effect.

	Index	
Method	*PESQ*	*STOI*
Deep neural networks (*DNN*) [55]	0.694	0.106
Recurrent deep stacking networks (*RDSN*) [55]	0.823	0.112
CNMF + JADE+ adaptive wavelet transform	0.305	0.106

small-sample, unsupervised speech separation in complex scenarios. The speech separation algorithm proposed in this study can satisfy the needs of small sample and unsupervised speech separation. In addition, the total computational overhead of the shallow model is smaller than that of the deep model. As opposed to the deep model, the shallow model is more suitable for application scenarios with high real-time requirements. Given the comparison of the two models synthetically, the algorithm proposed in this study is considered to be more suitable for target speaker speech extraction in the complex multispeaker scenario.

5. Conclusion

The development of IoT technology promotes the rapid development of intelligent voice systems, and the efficient processing of signal data acquired by speech sensors becomes imminent. Thus, an unsupervised speech separation algorithm based on the combination of CNMF and JADE is proposed in this study. Through simulation experiments, it is well demonstrated that the proposed algorithm can effectively separate the target speech signals contained in the mixed speech signals. In addition, for the separated speech signal is weak and out of frame, this study also proposes an adaptive wavelet transform method to enhance the separated speech signal. As revealed from the results, the proposed algorithm in this study can enhance the separated speech signals. The comprehensive experimental results can prove that the proposed algorithm is very competitive in the processing of single-channel mixed speech separation problem. The algorithm is highly versatile and robust, capable of technically supporting other researchers in processing highly noisy signal data collected by sensors.

Speech separation, especially single-channel speech separation, has been a hotspot and difficult research area. In addition, as IoT technology is being developed and applied, separating high-quality speech signals has become an urgent task. Speech signals exhibit obvious spatio-temporal structures and nonlinear relationships, and most of the conventional speech classification methods are shallow structures, and the mentioned results are more limited in their ability to tap into the mentioned nonlinear structural information. In recent years, as deep learning is advancing, it has been suggested that the nonlinear processing and feature learning capabilities of deep models exhibit obvious advantages in addressing speech separation problems. Moreover, some results of processing speech signals with deep learning have been published. As deep learning computing is leaping forward, deep models (e.g., DNN, DSN, CNN, RNN, Deep NMF, and LSTM) will definitely be more competitive in speech separation problems. In the future, the use of deep learning techniques in speech separation will definitely become a research hotspot.

Data Availability

We are using the TIMIT dataset, which can be found at https://academictorrents.com/details/34e2b78745138186976cbc27939b1b34d18bd5b3/techamp; hit =1amp; filelist =1.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61902232, 61902231), the Natural Science Foundation of Guangdong Province (2019A1515010943), the Key Project of Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Natural Science) (2018KZDXM035), the Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Special Projects in Artificial Intelligence) (2019KZDZX1030), and the 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (2020LKSFG04D).

References

[1] S. S. Hao, Y. P. Wang, and B. C. Lv, "A new optimisation model and algorithm for virtual optical networks," *International Journal of Sensor Networks*, vol. 29, no. 4, pp. 252–261, 2019.

[2] C. X. Ji, Y. P. Wang, Z. Q. Xu, and X. Li, "A new model and algorithm for RSA problem in elastic optical networks," *International Journal of Sensor Networks*, vol. 31, no. 3, pp. 145–155, 2019.

[3] M. Ye, Y. Wang, C. Dai, and X. Wang, "A hybrid genetic algorithm for the minimum exposure path problem of wireless sensor networks based on a numerical functional extreme model," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 8644–8657, 2016.

[4] H. Y. Liu, Y. P. Wang, and N. L. Fan, "A hybrid deep grouping algorithm for large scale global optimization," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 6, pp. 1112–1124, 2020.

[5] Y. Wang, H. Liu, F. Wei, T. Zong, and X. Li, "Cooperative coevolution with formula-based variable grouping for large-scale global optimization," *Evolutionary Computation*, vol. 26, no. 4, pp. 569–596, 2018.

[6] X. Jaureguiberry, E. Vincent, and G. Richard, "Fusion methods for speech enhancement and audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1266–1279, 2017.

[7] S. Srinivasan and D. L. Wang, "Robust speech recognition by integrating speech separation and hypothesis testing," *Speech Communication*, vol. 52, no. 1, pp. 72–81, 2010.

[8] L. Ferrer, M. K. Nandwana, M. Mclaren, D. Castan, and A. Lawson, "Toward fail-safe speaker recognition: trial-based calibration with a reject option," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 140–153, 2018.

[9] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and E. Moulines, "A blind source separation techinique based on second order statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 434–444, 2002.

[10] H. Zhang, G. Hua, L. Yu, Y. Cai, and G. Bi, "Underdetermined blind separation of overlapped speech mixtures in time-frequency domain with estimated number of sources," *Speech Communication*, vol. 89, pp. 1–16, 2017.

[11] B. Gao, W. L. Woo, and S. S. Dlay, "Adaptive sparsity non-negative matrix factorization for single-channel source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 989–1001, 2011.

[12] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised single-channel separation of nonstationary signals using gammatone filterbank and Itakura–Saito nonnegative matrix two-dimensional factorizations," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 3, pp. 662–675, 2013.

[13] N. Tengtrairat, W. L. Woo, S. S. Dlay, and B. Gao, "Online noisy single-channel source separation by adaptive spectrum amplitude estimator and masking," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1881–1895, 2016.

[14] Z. Fan, Y. Lai, and J. R. Jang, "SVSGAN: singing voice separation via generative adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 726–730, Calgary, AB, Canada, April 2018.

[15] M. Michelashvili, S. Benaim, and L. Wolf, "Semi-supervised monaural singing voice separation with a masking network trained on synthetic mixtures," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 291–295, Brighton, UK, May 2019.

[16] X. Sun, J. Xu, Y. Ma, T. Zhao, and S. Ou, "Single-channel blind source separation based on attentional generative adversarial network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, pp. 1–8, 2020.

[17] L. Xie, L. Tan, and M. W. Mak, "Guest editorial: advances in deep learning for speech processing," *Journal of Signal Processing Systems*, vol. 90, no. 7, pp. 1–3, 2018.

[18] T. Ogunfunmi, R. P. Ramachandran, R. Togneri, Y. Zhao, and X. Xia, "A primer on deep learning architectures and applications in speech processing," *Circuits, Systems, and Signal Processing*, vol. 38, no. 8, pp. 3406–3432, 2019.

[19] T. Zhao, Y. Zhao, and C. Xin, "Ensemble acoustic modeling for CD-DNN-HMM using random forests of phonetic decision trees," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 187–196, 2016.

[20] Z. Yan, H. Qiang, and X. Jian, "A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR," *Mathematics of Computation*, vol. 44, no. 170, pp. 519–521, 2013.

[21] Y. Wang, J. Du, L. R. Dai, and C. H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1535–1546, 2017.

[22] J. Du, Y. Tu, L.-R. Dai, and C. H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1424–1437, 2017.

[23] S. Behnke, "Discovering hierarchical speech features using convolutional non-negative matrix factorization," in *Proceedings of the International Joint Conference on Neural Networks, 2003*, Portland, OR, USA, July 2003.

[24] B. Chen, G. Polatkan, G. Sapiro, D. Blei, D. Dunson, and L. Carin, "Deep learning with hierarchical convolutional factor analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1887–1901, 2013.

[25] D. T. Pham and J. F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[26] C. Vaz, D. Dimitriadis, S. Thomas, and S. Narayanan, "CNMF-based acoustic features for noise-robust ASR," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, pp. 5735–5739, Shanghai, China, March 2016.

[27] Y. Zhang, S. Qi, and L. Zhou, "Single channel blind source separation for wind turbine aeroacoustics signals based on variational mode decomposition," *IEEE Access*, vol. 6, pp. 73952–73964, 2018.

[28] Z. Yan, S. Sheng-kai, L. Yue, and W. Jia-qi, "Sonar echo signal processing based on Convolution Blind source separation," in *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 130–134, Shanghai, China, September 2020.

[29] B. Wiem, B. M. Mohamed Anouar, P. Mowlaee, and B. Aicha, "Unsupervised single channel speech separation based on optimized subspace separation," *Speech Communication*, vol. 96, pp. 93–101, 2017.

[30] S. Mavaddaty, S. M. Ahadi, and S. Seyedin, "A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation," *Speech Communication*, vol. 76, pp. 42–60, 2015.

[31] Q. H. Lin, F. L. Yin, T. M. Mei, and H. Liang, "A blind source separation based method for speech encryption," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 6, pp. 1320–1328, 2006.

[32] H. T. Fan, J. W. Hung, X. Lu, S.-S. Wang, and Y. Tsao, "Speech enhancement using segmental nonnegative matrix factorization," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[33] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Transactions on Audio, Speech, & Language Processing*, vol. 26, no. 2, pp. 281–295, 2017.

[34] T. Pham, Y. S. Lee, Y. A. Chen, and J.-C. Wang, "A review on speech separation using NMF and its extensions," in *2015 International Conference on Orange Technologies (ICOT)*, pp. 26–29, Hong Kong, China, December 2016.

[35] S. Lee, D. Han, and H. Ko, "Single-channel speech enhancement method using reconstructive NMF with spectrotemporal speech presence probabilities," *Applied Acoustics*, vol. 117, pp. 257–262, 2017.

[36] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *2011 17th International Conference on Digital Signal Processing (DSP)*, pp. 1–6, Corfu, Greece, July 2011.

[37] T. Pham, Y. S. Lee, Y. B. Lin, T.-C. Tai, and J. Wang, "Single channel source separation using sparse NMF and graph regularization," in *ASE BD&SI '15: Proceedings of the ASE BigData & SocialInformatics 2015*, p. 55, New York, NY, USA, October 2015.

[38] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 2015.

[39] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: how models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[40] R. J. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech & Language*, vol. 24, no. 1, pp. 16–29, 2010.

[41] P. Mowlaee, R. Saeidi, M. G. Christensen et al., "A joint approach for single-channel speaker identification and speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2586–2601, 2012.

[42] R. Saeidi, P. Mowlaee, T. Kinnunen et al., "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, August 2010.

[43] Z. Jin and D. L. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1091–1102, 2011.

[44] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[45] H. Zhang, X. Zhang, S. Nie, G. Gao, and W. Liu, "A pairwise algorithm for pitch estimation and speech separation using deep stacking network," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 246–250, South Brisbane, QLD, Australia, April 2015.

[46] J. Chang and D. L. Wang, "Robust speaker recognition based on DNN/i-vectors and speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5415–5419, New Orleans, LA, USA, March 2017.

[47] J. L. Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70, South Brisbane, QLD, Australia, April 2015.

[48] S. Wisdom, T. Powers, J. Pitton, and L. Atlas, "Building recurrent networks by unfolding iterative thresholding for sequential sparse recovery," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, March 2017.

[49] F. Weninger, H. Erdogan, S. Watanabe et al., "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Latent Variable Analysis and Signal Separation. LVA/ICA 2015. Lecture Notes in Computer Science, vol 9237*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds., pp. 91–99, Springer, Cham, 2015.

[50] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[51] Y. H. Tu, J. Du, and C. H. Lee, "A speaker-dependent approach to single-channel joint speech separation and acoustic modeling based on deep neural networks for robust recognition of multi-talker speech," *Journal of Signal Processing Systems*, vol. 90, pp. 963–973, 2017.

[52] D. Yu, M. Kolbk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent

multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, New Orleans, LA, USA, March 2017.

[53] M. Kolbk, D. Yu, Z. H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[54] Y. Wang, J. Du, L. R. Dai, and C.-H. Lee, "A maximum likelihood approach to deep neural network based nonlinear spectral mapping for single-channel speech separation," in *Interspeech 2017*, pp. 1178–1182, Stockholm, Sweden, August 2017.

[55] Z. Q. Wang and D. L. Wang, "Recurrent deep stacking networks for supervised speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 71–75, New Orleans, LA, USA, March 2017.

[56] N. Dey and A. S. Ashour, "Applied examples and applications of localization and tracking problem of multiple speech sources," in *Direction of Arrival Estimation and Localization of Multi-Speech Sources. SpringerBriefs in Electrical and Computer Engineering*, pp. 35–48, Springer, Cham, 2018.

[57] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, 2002.

[58] M. Joham, K. Kusume, M. H. Gzara, W. Utschick, and J. A. Nossek, "Transmit Wiener filter for the downlink of TDDDS-CDMA systems," in *IEEE Seventh International Symposium on Spread Spectrum Techniques and Applications*, pp. 9–13, Prague, Czech Republic, September 2002.

[59] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 191–195, Shanghai, China, March 2016.

[60] T. Gülzow, A. Engelsberg, and U. Heute, "Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement," *Signal Processing*, vol. 64, no. 1, pp. 5–19, 1998.

[61] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[62] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[63] R. Plomp, "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," *Journal of Speech and Hearing Research*, vol. 29, no. 2, pp. 146–154, 1986.

[64] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, Salt Lake City, UT, USA, May 2002.

[65] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA, March 2010.

[66] R. Takashima, Y. Kawaguchi, Q. Sun, T. Sumiyoshi, and M. Togami, "An application of noise-robust speech translation using asynchronous smart devices," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1592–1595, Kuala Lumpur, Malaysia, December 2017.