Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.030180
Article

Tech Science Press

Speech Separation Algorithm Using Gated Recurrent Network Based on Microphone Array

Xiaoyan Zhao[1,*], Lin Zhou[2], Yue Xie[1], Ying Tong[1] and Jingang Shi[3]

[1]School of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing, 211167, China
[2]School of Information Science and Engineering, Southeast University, Nanjing, 210096, China
[3]University of Oulu, Oulu, 90014, FI, Finland
*Corresponding Author: Xiaoyan Zhao. Email: xiaoyanzhao@njit.edu.cn
Received: 20 March 2022; Accepted: 07 January 2023

Abstract: Speech separation is an active research topic that plays an important role in numerous applications, such as speaker recognition, hearing prosthesis, and autonomous robots. Many algorithms have been put forward to improve separation performance. However, speech separation in reverberant noisy environment is still a challenging task. To address this, a novel speech separation algorithm using gate recurrent unit (GRU) network based on microphone array has been proposed in this paper. The main aim of the proposed algorithm is to improve the separation performance and reduce the computational cost. The proposed algorithm extracts the sub-band steered response power-phase transform (SRP-PHAT) weighted by gammatone filter as the speech separation feature due to its discriminative and robust spatial position information. Since the GRU network has the advantage of processing time series data with faster training speed and fewer training parameters, the GRU model is adopted to process the separation features of several sequential frames in the same sub-band to estimate the ideal Ratio Masking (IRM). The proposed algorithm decomposes the mixture signals into time-frequency (TF) units using gammatone filter bank in the frequency domain, and the target speech is reconstructed in the frequency domain by masking the mixture signal according to the estimated IRM. The operations of decomposing the mixture signal and reconstructing the target signal are completed in the frequency domain which can reduce the total computational cost. Experimental results demonstrate that the proposed algorithm realizes omnidirectional speech separation in noisy and reverberant environments, provides good performance in terms of speech quality and intelligibility, and has the generalization capacity to reverberate.

Keywords: Microphone array; speech separation; gate recurrent unit network; gammatone sub-band steered response power-phase transform spatial spectrum

1 Introduction

Speech separation is the task of separating target speech from interference signals. Speech separation is an important research topic in the field of speech signal processing with a wide range of applications, including teleconferencing, speaker recognition, hearing prosthesis, and autonomous robots [1–3]. For instance, teleconferencing systems need to separate the near-end target speech from interfering signals before transmitting to the far-end listeners. In real scenarios, the performance of automatic speech recognition systems will degrade due to noise and interfering signals. As a crucial preprocessing step in automatic speech recognition systems, speech separation can improve the performance of automatic speech recognition systems by separating the target speech from background interference to remove interference signals. Speech separation is meaningful for hearing aid design because hearing aid devices should separate the target speaker's utterance from competing sound sources. Speech separation technology can extract individual sound sources from their mixture, which is very necessary to provide autonomous robots with machine audition capabilities.

Over the past few decades, a number of approaches have been put forward for the task of speech separation. The approaches can be categorized into monaural and array-based methods according to the number of microphones. The microphone array-based methods introduce spatial domain processing besides time domain processing and frequency domain processing, which boosts separation performance. The existing microphone array-based speech separation algorithms include beamforming methods [4], independent component analysis (ICA) methods [5], compressed sensing (CS)-based methods [6] and computational auditory scene analysis (CASA) methods [7]. The core of CASA-based speech separation algorithms include two stages: segmentation and grouping. The mixture speech is decomposed into time-frequency (TF) units as auditory perception segments in the segmentation stage, and the TF units from the same target sound source are reconstructed into an auditory data stream in the grouping stage.

Recently, a variety of CASA-based algorithms regard speech separation as a supervised learning problem and focus on the following components: training targets, acoustic features, and learning machines [8]. The commonly used training targets include ideal binary masking (IBM) [9], ideal ratio masking (IRM) [10], complex IRM [11], spectral magnitude mask (SMM) [12], phase sensitive mask (PSM) [13] and so on. The acoustic features play an important role in supervised speech separation. The commonly involved monaural features include Mel-frequency cepstral coefficient (MFCC) [14], gammatone frequency cepstral coefficient (GFCC), amplitude modulation spectrum (AMS) and so on. The commonly involved spatial features of multi-channel include inter-aural time differences (ITD), inter-aural level difference (ILD), inter-aural phase difference (IPD), cross-correlation function (CCF), generalized cross correlation (GCC), time difference of arrival (TDOA) and so on. The spatial features fully exploit the corresponding information from microphone array signals [15], and have the advantage of independent of the speaker and the content of the speech signal. The traditional learning machines include Gaussian mixture model (GMM), support vector machines (SVM) and so on. With the development of artificial neural network, deep learning approaches have been introduced for the task of speech separation in recent years.

The related researches for the introduction of deep learning to speech separation are as follows. Jiang et al. [16] combined the ITD, ILD and GFCC as the input features, treated IBM as training target, and utilized deep neural networks (DNN) for binaural speech separation. The approaches in [17,18] predicted the IRM value through DNN model. ILD and IPD were jointed as input features in [17], and the combination of ITD, ILD and monaural feature was treated as input features in [18]. Zhou et al. [19] defined a modified IRM and trained the long short-term memory

(LSTM) network model for binaural speech separation with input features consisting of CCF, IID and ILD. The approach in [20] converted the combining features including logarithmic amplitude spectrum (LPS) and IPD function into high-dimensional vectors through Bi-directional short-term memory (BiLSTM), and utilized K-means clustering to classify the TF units. Zhao et al. [21] implemented speech separation and speaker recognition using deep recurrent neural network (DRNN). Venkatesan et al. [22] proposed an iterative-DNN-based speech separation algorithm to retrieve two concurrent speech signals in a room environment. The aforementioned binaural speech separation algorithms can realize speech separation when the speakers are located in the front half of the horizontal plane due to the symmetry of spatial information, while the microphone array provides multiple recordings and can realize omnidirectional speech separation. The approaches in [23–26] decomposed the mixture signal into TF units by short-time Fourier transform (STTF) and extracted the IPDs in STFT domain as features. In [23], the approach took the IPD function as the input feature of the deep neural network of U-net architecture, and calculated the masking value according to the direction of arrival (DOA) estimation. The approaches in [24–26] took the combination of IPD function and LPS as input features and adopted the LSTM network to estimate the mask value. LSTM is a special kind of recurrent neural network (RNN) which has advantages in processing time series speech signals.

Speech separation methods may face some challenges when used in enclosed environments, including the limitations of the aforementioned spatial features, the insufficient speech separation performance in reverberant noisy conditions, and the need to reduce computational cost for practical applications. The aforementioned spatial features such as IPD, ILD, ITD, TDOA, CCF and GCC have certain limitations. IPD has the problem of high-frequency wrapping. ILD is not suitable for ordinary omnidirectional microphone arrays for far filed case. CCF and GCC suffer from the drawback of lack of robustness to noise and reverberation, resulting in incorrect ITD and TDOA estimates. As our previous work described in [27,28], the steered response power-phase transform (SRP-PHAT) spatial power spectrum contains robust spatial location information and is independent of the content of the speech signal. Therefore, the sub-band SRP-PHAT weighted by gammatone bandpass filter is adopted as the speech separation cue in this paper. Speech separation performance in adverse acoustic scenarios is still far from perfect due to reverberation and noise. Spatial features between consecutive speech frames are correlated, and learning the temporal dynamics of spatial features using recurrent neural networks will help improve speech separation performance. LSTM provides a good performance in processing time series data [29]. However, the introduction of a lot training parameters leads to insufficient training efficiency of LSTM. The gate recurrent unit (GRU) network is a variant of LSTM with fewer training parameters [30,31]. The GRU network combines the input gate and the forget gate into one gate: the update gate. At the same time, the GRU network does not introduce additional memory units, while introduces a linear dependency between the current state and the historical state. Compared with LSTM, the GRU network can accelerate the training procedure and provide a comparable performance [32]. Therefore, we introduce GRU network to model temporal dynamics of spatial features. gammatone filter bank is often used to decompose mixture signal into TF units. The existing methods conduct the auditory segmentation by convolving the mixture signal with the impulse response of gammatone filter, and obtain the TF unit of the target speech signal by multiplying mixture signal and the masking value in the time domain [15,19]. A large number of multiplication operations are required in the speech segmentation and grouping stage, resulting in computational cost consuming. The algorithm proposed in this paper decomposes the mixture signal and separates target speech in the frequency domain without the time domain convolution operation, thereby reducing the total computational cost. In summary, the main contributions of this paper are:

- Feature extraction: The sub-band SRP-PHAT weighted by gammatone bandpass filter is extracted as the separation feature, which fully exploits the phase information of the microphone array signals. The SRP-PHAT spatial power spectra at different positions are much more discriminative and robust than GCCs or TDOAs.
- GRU Model for IRM estimation: Considering the temporal dynamics of spatial features, a GRU model for IRM estimation is presented, which has the advantages of fewer training parameters and faster training speed.
- Scheme for speech segmentation and grouping: Different from the existing algorithms using gammatone filter bank to decompose cochleagram in the time domain, the proposed algorithm presents a scheme for speech segmentation and grouping in the frequency domain so that the computational cost is reduced.

Through experimental evaluation, the proposed algorithm has been shown to provide good performance in terms of speech quality and intelligibility in noisy and reverberant environments and have the generalization capacity to reverberation.

The rest of this paper is organized as follows. Section 2 presents the overview of the proposed speech separation system. Section 3 describes the proposed speech separation algorithm based on GRU, including preprocess, feature extraction, speech separation and reconstruction, the architecture of GRU network and the training of GRU network. The experimental results and analysis are presented in Section 4. The conclusion is drawn in Section 5.

2 System Architecture

The core idea of our algorithm is to use GRU network to approximate the IRM value through spatial feature. The proposed algorithm treats the IRM estimation problem as a regression task. Fig. 1 illustrates the overall architecture of the proposed speech separation system. The microphone array-based speech separation system using GRU network includes two phases: the training phase and the testing phase. The system inputs are the mixture signals received by microphone array. The mixture signals are decomposed into TF units by gammatone filter bank in the frequency domain. Then, the gammatone-weighted sub-band SRP-PHAT spatial spectrum [28] is extracted in each TF unit as the spatial feature for speech separation. Furthermore, considering the temporal dynamics, the spatial features of several sequential frames in the same sub-band are concatenated to form a spatial feature matrix, which is treated as GRU network input for the central TF unit. In the training phase, the GRU networks of all sub-bands are trained to approximate the IRM targets through the spatial features. To improve the robustness and generalization of the system, training signals with diverse reverberation and noise are taken together to train the GRU networks. During the testing phase, the trained GRU network outputs estimated IRM values within each TF unit of the mixture testing signal. Then, the target speech in each TF unit is reconstructed in the frequency domain by masking the mixture signal according to the estimated IRM. Finally, the reconstructed target speech signals in the time domain are obtained by conducting inverse Fourier transform on the combination of the signals of all sub-band.

3 Speech Separation Algorithm Using GRU Network

3.1 Preprocess

The physical model for speech signal from multi-speakers received by mth microphone in indoor scenarios can be formulated as follows:

$$x_m(t) = h_{m,1}(t) * s_1(t) + h_{m,2}(t) * s_2(t) + v_m(t) \quad m = 1, \ldots M \tag{1}$$

where $x_m(t)$ represents the mixture signal received by the mth microphone, $s_1(t)$ and $s_2(t)$ denote two speech source signals, $h_{m,1}(t)$ and $h_{m,2}(t)$ represent the room impulse responses from speech sources to the mth microphone, "$*$" denotes the linear convolution, $v_m(t)$ is additive noise for the mth microphone, and M is the number of microphones.

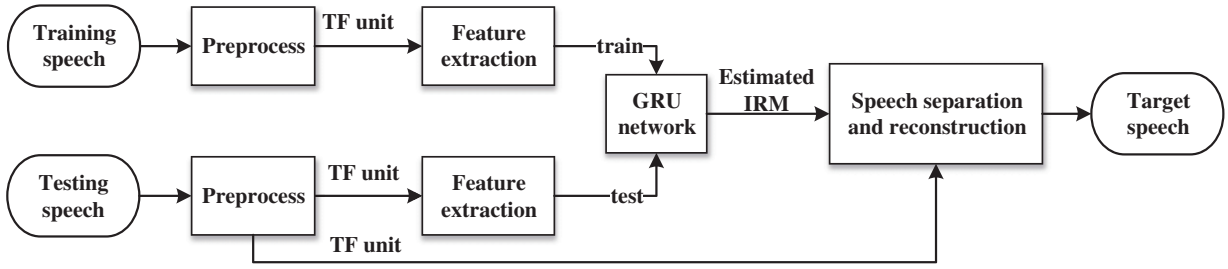


Figure 1: Overall architecture of microphone array-based speech separation system using GRU network

Short-time Fourier transform (STFT) is conducted on the microphone array signal after framing and windowing:

$$X_m(k,\omega) = \int_0^T x_m(k,t)\, e^{-j\omega t}\ m = 1 \ldots M \tag{2}$$

where $X_m(k,\omega)$ represents the spectrum of the mth microphone signal at frame k, ω is the frequency, and T is the frame length.

Gammatone filter bank is used to simulate the time-frequency analysis to acoustic signals. The impulse response of the ith gammatone filter is defined as:

$$g_i(t) = ct^{n-1}e^{-2\pi b_i t}\cos(2\pi f_i t + \varphi), t > 0 \tag{3}$$

where c denotes the gain coefficient, n denotes the filter order, b_i denotes the decay coefficient, f_i denotes the central frequency of the ith filter, and φ denotes the phase. The central frequencies of gammatone filters range from 50 to 8000 Hz on the equivalent rectangular bandwidth (ERB). Unlike the existing algorithms, the proposed algorithm decomposes the microphone array signal into TF units in the frequency domain as follows:

$$X_m(k,i,\omega) = X_m(k,\omega)\, G_i(\omega) \tag{4}$$

where $X_m(k,i,\omega)$ represents the spectrum of the mth microphone signal at frame k and sub-band i, $X_m(k,i,\omega)$ is the TF unit for speech separation, and $G_i(\omega)$ is the Fourier transforms of $g_i(t)$. In this paper, the number of sub-band is 32.

3.2 Feature Extraction

The spatial position of the speakers is sparse, and the spatial position information is independent of the speaker and the content of the speech signal. The mixture signal can be effectively separated according to the spatial position information without establishing a statistical model of the source signal parameters. Furthermore, the spatial information has the abilities of simultaneous organization and sequential organization. Therefore, the spatial position information extracted from the array signals is used as the feature for speech separation.

As our previous work described in [27,28], the SRP-PHAT spatial power spectrum of the array signals contains spatial position information. The SRP-PHAT function fully exploits the phase information of the microphone array signals. The SRP-PHAT spatial power spectra at different positions are much more discriminative and robust than GCCs or TDOAs. Therefore the SRP-PHAT spatial power spectrum is exploited as the separation feature. In each TF unit, the extracted feature is the sub-band SRP-PHAT weighted by gammatone bandpass filter, and it is defined as:

$$P(k,i,\mathbf{r}) = \sum_{m=1}^{M}\sum_{n=m+1}^{M}\int_{-\infty}^{\infty}|G_i(\omega)|^{\gamma}\frac{X_m(k,i,\omega)\,X_n^*(k,i,\omega)}{\left|X_m(k,i,\omega)\,X_n^*(k,i,\omega)\right|}e^{j\omega\Delta\tau_{mn}(\mathbf{r})}d\omega$$
$$= \sum_{m=1}^{M}\sum_{n=m+1}^{M}\int_{-\infty}^{\infty}|G_i(\omega)|^{\gamma}\frac{X_m(k,\omega)\,X_n^*(k,\omega)}{\left|X_m(k,\omega)\,X_n^*(k,\omega)\right|}e^{j\omega\Delta\tau_{mn}(\mathbf{r})}d\omega \tag{5}$$

where $P(k,i,\mathbf{r})$ represents the response power of TF unit at frame k and sub-band i, $\mathbf{r}$ is the steering position, $\Delta\tau_{mn}(\mathbf{r})$ is the propagation delay difference from steering position $\mathbf{r}$ to the mth microphone and the nth microphone, $\Delta\tau_{mn}(\mathbf{r})$ is only related to the azimuth of the steering position $\mathbf{r}$ in the far-filed case, γ denotes the weighting order, and γ is set to 1 in this paper. From Eq. (5), we note that the phase information of frequency components in each TF unit is weighted by gammatone bandpass filter.

The Gammatone-weighted sub-band SRP-PHATs within a TF unit are arranged into a vector, which can be expressed as follows:

$$\boldsymbol{P}(k,i) = \left[P(k,i,\mathbf{r}_1),\quad P(k,i,\mathbf{r}_2),\quad \ldots\quad P(k,i,\mathbf{r}_D)\right] \tag{6}$$

where $\boldsymbol{P}(k,i)$ represents the spatial feature vector extracted in each TF unit, D is the number of steering positions. In the far-filed case, the argument $\mathbf{r}_d$ is simplified to the azimuth. The azimuth of steering position ranges from 0° to 360° with a step of 5°, corresponding to 72 steering positions. Thus the dimension of gammatone-weighted sub-band SRP-PHAT feature vector is 72. The gammatone-weighted sub-band SRP-PHAT feature vectors of different directions have good discrimination.

Due to the temporal dynamics of speech signals, the spatial feature of consecutive frames are highly correlated. Research shows that the inter-frame correlation of the spatial features is small when the frame interval exceeds 4. Thus the spatial features of 9 sequential frames (4 before and 4 after the current TF unit) in the same sub-band are concatenated to form the GRU input matrix. Therefore the dimension of GRU input matrix is 9×72.

3.3 Speech Separation and Reconstruction

The two speech source signals and noise are considered to be uncorrelated with each other within a given TF unit. IRM is used as the mask for recovering the target signal form the mixture signal received by microphone array. IRMs for speech source signals and noise are defined as follows [19]:

$$IRM_1(k,i) = \sqrt{\frac{E_1(k,i)^2}{E_1(k,i)^2 + E_2(k,i)^2 + E_v(k,i)^2}} \tag{7}$$

$$IRM_2(k,i) = \sqrt{\frac{E_2(k,i)^2}{E_1(k,i)^2 + E_2(k,i)^2 + E_v(k,i)^2}} \tag{8}$$

$$IRM_v(k,i) = \sqrt{\frac{E_v(k,i)^2}{E_1(k,i)^2 + E_2(k,i)^2 + E_v(k,i)^2}} \tag{9}$$

where $IRM_1(k,i)$ and $IRM_2(k,i)$ indicate the ideal masks of the target signals in TF unit at frame k and sub-band i, $IRM_v(k,i)$ indicates the mask of the noise, $E_1(k,i)^2$ and $E_2(k,i)^2$ represent the energy of two speech signals within the TF unit respectively, and $E_v(k,i)^2$ represents the energy of additive noise within the TF unit.

During the testing phase, the output of GRU network is an estimation of the ideal mask. The target speech in each TF unit is reconstructed by masking the mixture signal according to the estimated IRM as follows:

$$\widetilde{S}_1(k,i,\omega) = \widetilde{I}RM_1(k,i)\, X(k,i,\omega) \tag{10}$$

$$\widetilde{S}_2(k,i,\omega) = \widetilde{I}RM_2(k,i)\, X(k,i,\omega) \tag{11}$$

where $\widetilde{S}_1(k,i,\omega)$ and $\widetilde{S}_2(k,i,\omega)$ represent the reconstructed spectrums of two target speech signals in TF unit at frame k and sub-band i, $\widetilde{I}RM_1(k,i)$ and $\widetilde{I}RM_2(k,i)$ are the estimated IRM values from GRU network for two speakers within a given TF unit, $X(k,i,\omega)$ is the spectrum of mixture signal from any one of the microphones. Thereafter, for a given TF unit, the target signal in the time domain is calculated as follows:

$$\widetilde{s}_1(k,i,t) = IDFT\left(\frac{\widetilde{S}_1(k,i,\omega)}{G_i(\omega)}\right) \omega \in \left[\omega_i^L, \omega_i^H\right] \tag{12}$$

$$\widetilde{s}_2(k,i,t) = IDFT\left(\frac{\widetilde{S}_2(k,i,\omega)}{G_i(\omega)}\right) \omega \in \left[\omega_i^L, \omega_i^H\right] \tag{13}$$

where ω_i^L and ω_i^H denote the lower and upper bounds of the ith sub-band respectively, and $IDFT(\cdot)$ denotes the inverse Fourier transform operation.

Subsequently, the target signal in a given frame is recovered by combining the signals of all sub-bands as follows:

$$\widetilde{s}_1(k,t) = \sum_i \widetilde{s}_1(k,i,t) \tag{14}$$

$$\widetilde{s}_2(k,t) = \sum_i \widetilde{s}_2(k,i,t) \tag{15}$$

where $\widetilde{s}_1(k,t)$ and $\widetilde{s}_2(k,t)$ represent the time domain signals of two target sources at frame k. Finally, the reconstructed target speech signals are obtained by combining the frame signals. Substituting Eqs. (12) and (13) into Eqs. (14) and (15) respectively and considering the properties of the inverse Fourier transform, Eqs. (14) and (15) can be rewritten as follow:

$$\widetilde{s}_1(k,t) = \sum_i IDFT\left(\frac{\widetilde{S}_1(k,i,\omega)}{G_i(\omega)}\right) = IDFT\left(\sum_i \frac{\widetilde{S}_1(k,i,\omega)}{G_i(\omega)}\right) \omega \in \left[\omega_i^L, \omega_i^H\right] \tag{16}$$

$$\widetilde{s}_2(k,t) = \sum_i IDFT\left(\frac{\widetilde{S}_2(k,i,\omega)}{G_i(\omega)}\right) = IDFT\left(\sum_i \frac{\widetilde{S}_2(k,i,\omega)}{G_i(\omega)}\right) \omega \in \left[\omega_i^L, \omega_i^H\right] \tag{17}$$

In practical applications, fast Fourier transform (FFT) is usually used instead of STFT to process the signal. Gammatone filter bank can be designed and stored in advance. In the segmentation stage, the proposed method decomposes mixture speech into TF units according to Eq. (4). The FFT operation on the mixture speech needs to perform $(L\log_2 L)/2$ complex multiplications, where L is the length of FFT. Multiplying $X_m(k,i,\omega)$ and $G_i(\omega)$ needs to perform L complex multiplication.

Hence, a total of $(L\log_2 L)/2 + LI$ complex multiplications are required to decompose a frame of signal, where I is the number of sub-band. However, the methods in [15,19] conduct the auditory segmentation by convolving the mixture speech with the impulse response of gammatone filter in the time domain. The convolution can be realized by overlap-add method to reduce the computational cost. It is necessary to perform FFT on mixture speech to obtain $X_m(k,i,\omega)$, and perform inverse fast Fourier transform (IFFT) on the product of $X_m(k,i,\omega)$ and $G_i(\omega)$. Hence, it requires $(L\log_2 L + L)\cdot I$ complex multiplications in total for segmentation in [15,19]. In the grouping stage, the computational cost of reconstructing the target speech in each TF unit according to Eqs. (10) and (11) is similar to the computational cost in [15,19]. To recover the auditory data stream of the target speech, the proposed method needs $(L\log_2 L)/2 + LI$ complex multiplications according to Eqs. (16) and (17), while the methods in [15,19] need $(L\log_2 L + L)\cdot I$ complex multiplications. Therefore, the proposed scheme for speech segmentation and grouping requires fewer complex multiplications and reduces the computational cost effectively.

3.4 The Architecture of GRU Network

Speech signal has characteristics of temporal dynamics, and RNN has advantages in processing time series data. Although both GRU and LSTM are recurrent networks, the GRU network has fewer training parameters and accelerated training procedure, and provides comparable performance. Therefore, RNN with GRU is adopted to learn the temporal dynamics of spatial features for speech separation. As depicted in Fig. 2, the architecture of the network includes an input layer, two GRU layers, two fully connected layers, and an output layer. Since the time-step is set to 9 (4 before and 4 after the current TF unit), the data of input layer is the spatial feature matrix of size 9×72 which is described in Section 3.2. The input layer is followed by two GRU layers. Each GRU layer contains 256 bidirectional GRU units. The two GRU layers fully encode the information of the input signal in the time-step. The last GRU layer is followed by two fully connected layers. For each fully connected layer, batch normalized (BN) and rectified linear unit (ReLU) activation function are performed after linear operation. The dropout method is introduced in each GRU layer and fully connected layer to prevent overfitting. The output layer outputs the estimated IRM for each speaker with different directions. The azimuth of speaker ranges from 0° to 360° with a step of 10°, corresponding to 36 training directions. Therefore, the output layer contains 37 neurons, corresponding to 36 directions and noise. For the output layer, the softmax regression model is adopted to convert the feature data into the IRM value of the target signal at a given direction.

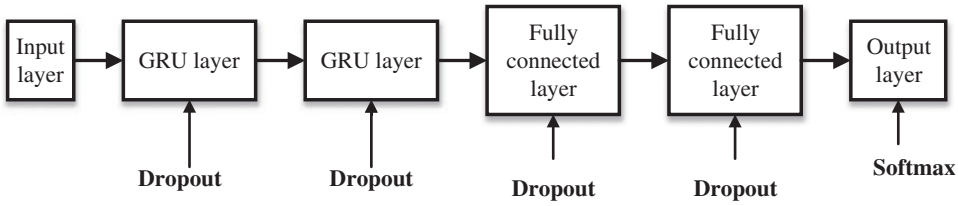


Figure 2: The GRU network architecture of the proposed algorithm

3.5 The Training of GRU

The speech separation system trains the models separately for 32 sub-bands. To improve the robustness and generality of model, the training data with diverse noisy and reverberant are used

together to train the GRU model. The input data of the GRU model is the sub-band SRP-PHAT spatial feature constructed in Section 3.2.

For each sub-band, the training target of GRU model is the label vector, which is expressed as follows:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 & y_2 \ldots & y_{36} & y_{37} \end{bmatrix} = \begin{bmatrix} 0, \ldots & IRM_1, \ldots & IRM_2, \ldots & IRM_v \end{bmatrix} \tag{18}$$

In the label vector, the values of the two elements whose positions in the vector are corresponding to the speakers' directions are IRM_1 and IRM_2, the value of the last element is IRM_v for noise, and the values of the other elements in the label vector are 0. The IRM values are calculated by Eqs. (7)–(9). From Eq. (18), we note that the label vector is sparse.

The training of GRU model includes forward propagation process and back propagation process. In the forward propagation process, the features are transferred layer by layer, and the expression of the output layer is as follows:

$$\boldsymbol{Z} = \text{Softmax}\left(f\left(\boldsymbol{P}\right)\right) \tag{19}$$

where $\boldsymbol{Z}$ is the output of the GRU network; $\boldsymbol{P}$ is the input signal, that is sub-band SRP-PHAT spatial feature in this paper; and $f(\cdot)$ is total operation of the GRU network under the current model parameters, including gated loop unit operation of GRU, linear operation, BN operation, activation operation and so on.

The model parameters are updated through the back propagation algorithm. The mean square error (MSE) between the output of the GRU network and the training label vector is used as the loss function, which is expressed as follows:

$$J = \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{Z}||_2^2 \tag{20}$$

where $||\cdot||_2$ represents $L2$ norm.

The MSE loss function is minimized in the back propagation process to update the model parameters. The Adam optimizer is adopted. The number of training epochs is set to 50. The gradient decay factor is set to 0.9, the square gradient decay factor is set to 0.99, the initial learning rate is 0.001, and the mini-batch is set to 200. To prevent over-fitting in the training phase, the 7:3 cross validation is adopted and the dropout method with a ratio of 0.5 is introduced.

4 Simulation and Result Analysis

4.1 Simulation Setup

Simulation experiments are implemented to evaluate the performance of the proposed method. The dimensions of the simulated room are given as 7 m × 7 m × 3 m. A uniform circular array consisted of six omnidirectional microphones is located at (3.5, 3.5, 1.6 m) in the room. The radius of the array is 10 cm. The clean speech signals are selected randomly from the TIMIT database, with a sampling rate of 16 kHz. The room impulse response from speaker to microphone is generated by the Image method [33]. The reverberant signal is derived by convolving the clean speech signal with the room impulse response. The scaled Gaussian white noise is added to the sum of reverberant signals of two speech sources to generate the mixture signal received by microphone.

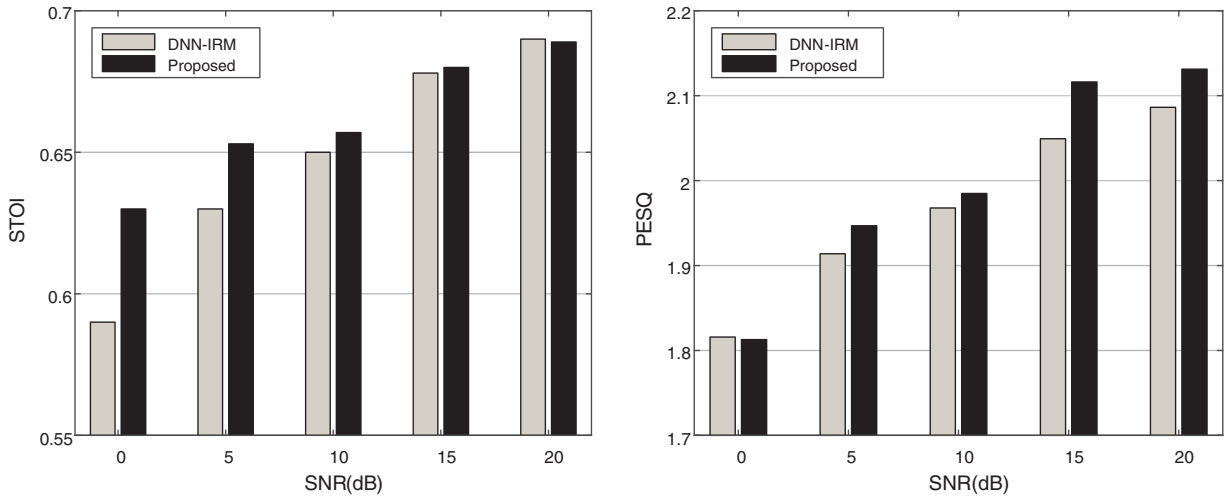
The speaker is in the far-field, and the azimuth varies between [0°, 360°] with a step of 10°. During the training phase, the SNR is varied from 0 to 20 dB with a step of 5 dB, and the reverberation time T60 is set to two levels as 0.2 and 0.6 s. The training data with different reverberation and noise are

taken together to train the GRU network. The microphone signals are segmented into 32-ms frame length with a shift of 16 ms.

The separation performance is measured by short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ). STOI is the metric of speech intelligibility, and PSEQ is the metric of speech quality, which is used to evaluate the auditory perception characteristics of speech. The performance of the proposed algorithm is compared with DNN-IRM method mentioned in [15].

4.2 Evaluation in Setup-Matched Environments

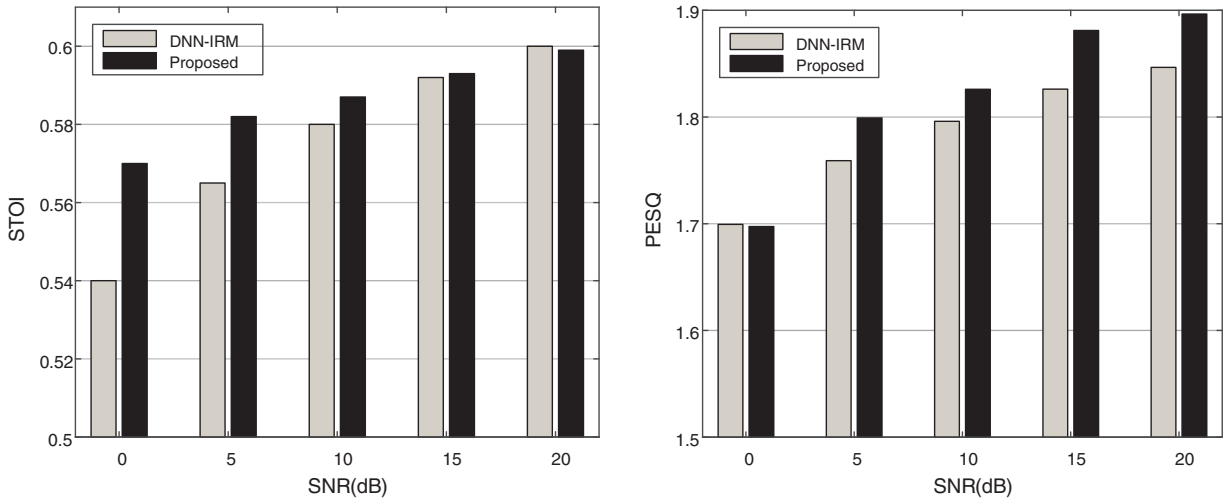
In this section, the speech separation performance is evaluated in the setup-matched environments, that is, the test signals and the training signals are generated under the same setup conditions. Figs. 3 and 4 depict the speech separation performance as a function of SNR for different algorithms under reverberation environments with T60 = 0.2 and T60 = 0.6 s, respectively.



(a) STOI comparison (b) PESQ comparison

Figure 3: Performance comparison of two algorithms with T60 = 0.2 s

From Figs. 3 and 4, it can be seen that the performance of speech separation algorithm deteriorates as the SNR decreases. Based on Figs. 3 and 4, we have found that the proposed algorithm achieves better speech intelligibility and quality than the DNN-IRM method in noisy and reverberant environments. The reason is that the proposed algorithm adopts the gammatone sub-band SRP-PHAT containing robust spatial position information as spatial feature, which effectively exploits the phase information of the microphone array signals and has good discrimination. And meanwhile the GRU network used by the proposed algorithm can introduce the temporal context. Furthermore, in terms of speech intelligibility, the STOI improvement of the proposed method compared with the DNN-IRM method is significant under low SNR environments (below 5 dB), that is about 0.02∼0.04 which means that the proposed method can significantly improve speech intelligibility in high noisy environments. In terms of speech quality, the proposed method and the DNN-IRM method have similar PESQ values in low SNR condition, and the proposed method presents higher PESQ values than those of the DNN-IRM method in moderate to high SNR condition with about 0.02∼0.07 improvement, which means that the proposed method can significantly improve speech quality in in moderate to high SNR condition.

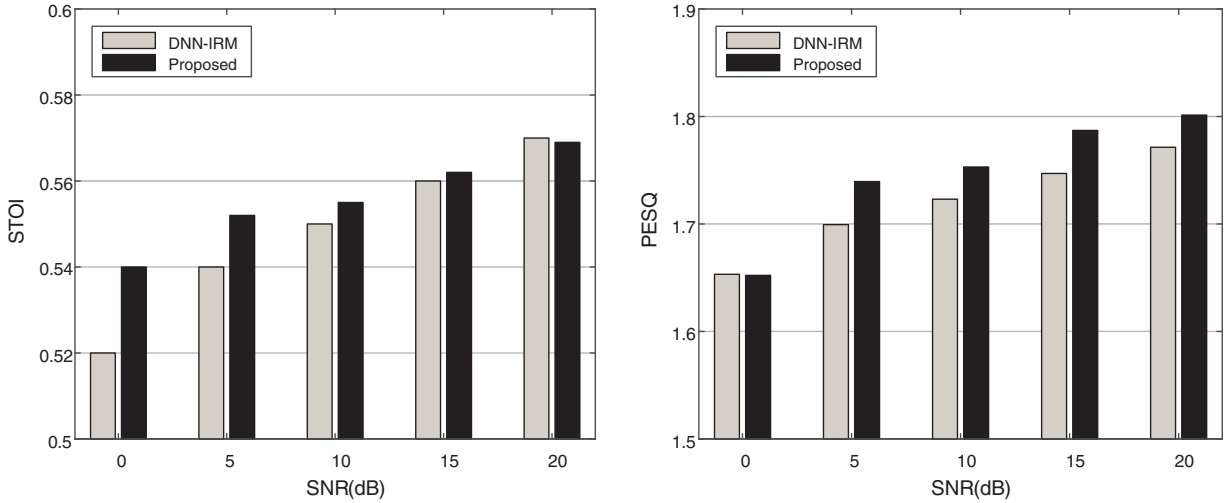


(a) STOI comparison (b) PESQ comparison

Figure 4: Performance comparison of two algorithms with T60 = 0.6 s

4.3 Evaluation in Setup-Unmatched Environments

In this section, we investigate the reverberation generalization of the proposed algorithm in untrained environments. For the testing signals, the reverberation time T60 is set to 0.8 s which is different from that of training signals. Fig. 5 depicts the performance comparison results.



(a) STOI comparison (b) PESQ comparison

Figure 5: Performance comparison of two algorithms with T60 = 0.8 s

Based on the results in Fig. 5, we have found that the proposed algorithm is superior to the DNN-IRM method in terms of speech intelligibility and quality. Compared to Fig. 4 with T60 = 0.6 s, the STOI and PESQ are slightly reduced, while the regularity of data variation is similar to those described in Section 4.2. Specifically, the STOI improvement increases from 0 to 0.02 as the SNR decreases from 20 to 0 dB. The proposed method and the DNN-IRM method have similar PESQ values in low

SNR condition, and the proposed method presents higher PESQ values than those of the DNN-IRM method in moderate to high SNR condition with about 0.03 improvement. Therefore, the proposed method achieves better speech intelligibility and quality in the untrained reverberant environments, demonstrating that the proposed method has better generalization performance to reverberate.

5 Conclusion

In this work, a speech separation algorithm using GRU network based on microphone array has been presented. Different from the existing algorithms using gammatone filter bank to decompose cochleagram in the time domain, the proposed algorithm decomposes the mixture speech and separates target speech in the frequency domain. The proposed scheme for speech segmentation and grouping can reduce the total computational cost for speech separation. The gammatone sub-band SRP-PHAT spatial power spectrum which contains robust spatial position information is exploited as the feature for speech separation. Considering the temporal dynamics of spatial features, the GRU network model which has the advantages of fewer training parameters and faster training speed, is adopted to estimate the IRM value for each target source in the TF unit. Experimental results show that the proposed algorithm can achieve omnidirectional speech separation, provide better speech quality and intelligibility both in the trained and untrained environments compared with DNN-IRM method, and have the generalization capacity to reverberation. The limitations of the proposed method in practical applications are: First, it is necessary to collect a sufficient number of training data covering variabilities including the speaker' positions, reverberation, and noises. Second, it is necessary to annotate the training data with ground-truth labels, which are not the direct information contained in mixture speech and need to be calculated according to Eqs. (7)–(9). In future work, we will adopt the transfer learning method to address the above issues.

Funding Statement: This work is supported by Nanjing Institute of Technology (NIT) fund for Research Startup Projects of Introduced talents under Grant No. YKJ202019, Nature Science Research Project of Higher Education Institutions in Jiangsu Province under Grant No. 21KJB510018, National Nature Science Foundation of China (NSFC) under Grant No. 62001215, and NIT fund for Doctoral Research Projects under Grant No. ZKJ2020003.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Bao, Q. Shao, X. Zhang, J. Jiang, Y. Xie *et al.,* "A novel system for recognizing recording devices from recorded speech signals," *Computers, Materials & Continua*, vol. 65, no. 3, pp. 2557–2570, 2020.

[2] J. Jo, H. Kim, I. Park, C. B. Jung and H. Yoo, "Modified viterbi scoring for HMM-based speech recognition," *Intelligent Automation & Soft Computing*, vol. 25, no. 2, pp. 351–358, 2019.

[3] J. Park and S. Kim, "Noise cancellation based on voice activity detection using spectral variation for speech recognition in smart home devices," *Intelligent Automation & Soft Computing*, vol. 26, no. 1, pp. 149–159, 2020.

[4] D. P. Jarrett, E. A. Habets and P. A. Naylor, "Signal-independent array processing," in *Theory and Applications of Spherical Microphone Array Processing*, vol. 9. New York, USA: Springer, pp. 93–111, 2017.

[5] H. Sawada, S. Araki, R. Mukai and S. Makino, "Blind extraction of dominant target sources using ICA and time-frequency masking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2165–2173, 2006.

[6] J. Su, H. Tao, M. Tao, D. Wang and J. Xie, "Narrow-band interference suppression via RPCA-based signal separation in time–frequency domain," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 11, pp. 5016–5025, 2017.
[7] D. L. Wang and G. J. Brown, "Fundamentals of Computational Auditory Scene Analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, vol. 1. Hoboken, NJ, USA: Wiley-IEEE Press, pp. 1–44, 2006.
[8] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
[9] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.
[10] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, pp. 7092–7096, 2013.
[11] D. S. Williamson, Y. Wang and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.
[12] Y. Wang, A. Narayanan and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.
[13] H. Erdogan, J. R. Hershey, S. Watanabe and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, South Brisbane, QLD, Australia, pp. 708–712, 2015.
[14] Y. Wang, K. Han and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.
[15] L. Zhou, Y. Xu, T. Y. Wang, K. Feng and J. G. Shi, "Microphone array speech separation algorithm based on TC-ResNet," *Computers, Materials & Continua*, vol. 69, no. 2, pp. 2705–2716, 2021.
[16] Y. Jiang, D. Wang, R. Liu and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.
[17] Y. Yu, W. Wang and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2016, no. 7, pp. 1–18, 2016.
[18] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.
[19] L. Zhou, S. Lu, Q. Zhong, Y. Chen, Y. Tang *et al.,* "Binaural speech separation algorithm based on long and short time memory networks," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1373–1386, 2020.
[20] L. Zhou, Y. Xu, T. Y. Wang, K. Feng and J. G. Shi, "Binaural speech separation algorithm based on deep clustering," *Intelligent Automation & Soft Computing*, vol. 30, no. 2, pp. 527–537, 2021.
[21] Y. Zhao, Z. Wang and D. Wang, "A Two-stage algorithm for noisy and reverberant speech enhancement," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, New Orleans, LA, pp. 5580–5584, 2017.
[22] R. Venkatesan and A. B. Ganesh, "Binaural classification-based speech segregation and robust speaker recognition system," *Circuits Systems & Signal Processing*, vol. 37, no. 9, pp. 1–29, 2017.
[23] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger and S. Gannot, "Multi-microphone speaker separation based on deep DOA estimation," in *27th European Signal Processing Conf.*, A Coruna, Spain, pp. 1–5, 2019.
[24] L. Chen, M. Yu, D. Su and D. Yu, "Multi-band pit and model integration for improved multi-channel speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Brighton, UK, pp. 705–709, 2019.
[25] Z. Q. Wang, J. Le Roux and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, pp. 1–5, 2018.

[26] Z. Q. Wang and D. L. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2019.
[27] X. Y. Zhao, S. W. Chen and L. Zhou, "Sound source localization based on SRP-PHAT spatial spectrum and deep neural network," *Computers, Materials & Continua*, vol. 64, no. 1, pp. 253–271, 2020.
[28] X. Y. Zhao, L. Zhou, Y. Tong, Y. Qi and J. G. Shi, "Robust sound source localization using convolutional neural network based on microphone array," *Intelligent Automation & Soft Computing*, vol. 30, no. 1, pp. 361–371, 2021.
[29] Mustaqeem, M. Ishaq and S. Kwon, "Short-term energy forecasting framework using an ensemble deep learning approach," *IEEE Access*, vol. 9, pp. 94262–94271, 2021.
[30] Mustaqeem and S. Kwon, "1d-cnn: Speech emotion recognition system using a stacked network with dilated cnn features," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 4039–4059, 2021.
[31] B. M, M. Swain, and Mustaqeem, "Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and Bi-GRU features," *Electronics*, vol. 11, no. 9, pp. 1–18, 2022.
[32] K. Cho, B. V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares *et al.,* "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *2014 Conf. on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1724–1734, 2014.
[33] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.