Binaural Speech Separation Algorithm Based on Long and Short Time Memory Networks

Lin Zhou[1, *], Siyuan Lu[1], Qiuyue Zhong[1], Ying Chen[1, 2], Yibin Tang[3] and Yan Zhou[3]

Abstract: Speaker separation in complex acoustic environment is one of challenging tasks in speech separation. In practice, speakers are very often unmoving or moving slowly in normal communication. In this case, the spatial features among the consecutive speech frames become highly correlated such that it is helpful for speaker separation by providing additional spatial information. To fully exploit this information, we design a separation system on Recurrent Neural Network (RNN) with long short-term memory (LSTM) which effectively learns the temporal dynamics of spatial features. In detail, a LSTM-based speaker separation algorithm is proposed to extract the spatial features in each time-frequency (TF) unit and form the corresponding feature vector. Then, we treat speaker separation as a supervised learning problem, where a modified ideal ratio mask (IRM) is defined as the training function during LSTM learning. Simulations show that the proposed system achieves attractive separation performance in noisy and reverberant environments. Specifically, during the untrained acoustic test with limited priors, e.g., unmatched signal to noise ratio (SNR) and reverberation, the proposed LSTM based algorithm can still outperforms the existing DNN based method in the measures of PESQ and STOI. It indicates our method is more robust in untrained conditions.

Keywords: Binaural speech separation, long and short time memory networks, feature vectors, ideal ratio mask.

1 Introduction

Speech separation focuses on separating target speech from interference, i.e., background noise, reverberation and interfering speech. As a front-to-end of speech signal processing system, it is widely used in various scenarios, e.g., smart homes, hearing aids, and speech interaction system.

In terms of the number of used microphones, speech separation methods can be divided into two categories such as monaural and array-based ones. In monaural methods, they

[1] School of Information Science and Engineering, Southeast University, Nanjing, 210096, China.

[2] Department of Psychiatry, Columbia University and NYSPI, New York, 10032, USA.

[3] College of Internet of Things Engineering, Hohai University, Changzhou, 213022, China.

[*] Corresponding Author: Lin Zhou. Email: Linzhou@seu.edu.cn.

Received: 15 February 2020; Accepted: 28 February 2020.

CMC. doi:10.32604/cmc.2020.010182

employ a variety of features, e.g., pitch [Han and Wang (2012)], Gammatone Frequency Cepstral Coefficient (GFCC) [Shao and Wang (2008)], and Delta Spectral Cepstral Coefficient (DSCC) [Kumar, Kim and Stern (2011)] to recognize target speech. However, the array-based methods can use more information from spatial configuration of sound source and acoustic environment to achieve better separation performance.

To our knowledge, the array-based speech separation is founded on the array signal. Here, three well-developed array signal processing methods, i.e., beamforming [Jarrett, Habets and Naylor (2017); Benesty, Chen and Huang (2008); DiBiase, Silverman and Brandstein (2001)], Independent Component Analysis (ICA) [Sawada, Araki, Mukai et al. (2006)], and Compressed Sensing (CS) [Su, Tao, Tao et al. (2017)], are briefly introduced. Beamforming aims to boost the signal arriving from a particular direction and attenuate interference from other directions. In general, noise attenuation depends on the size and configuration of the array. As mentioned in Wang et al. [Wang and Chen (2018)], Wang point out that the room reverberation reduces the utility of beamforming. Also, when the target signal and inference signal are too close to each other, beamforming can’t be performed due to its resolution limitation. As for ICA, it separates the target signal by the search of statistically independent and non-Gaussian components in multichannel signals. However, the existing ICA-based separation performance is unsatisfied in the reverberant environment, while its source permutation problem after separation is still unsolved. The CS based method provides another way to tackle the separation problem. It employs the sparsity of the speech signal in the time-frequency domain. The separation is achieved by the sparse representation of source signal from sampled signals. CS uses the dictionary to reconstruct the speech source. For example, the training and testing signal are both derived from the same speakers or the same speech content.

Nowadays, speech separation is frequently treated as a supervised learning problem. In the Computational Auditory Scene Analysis (CASA), the learning goal [Wang (2008); Wang, Brown and Darwin (2008)] is to compute Ideal Binary Mask (IBM) for each TF unit. Reports show [Sinex (2013)] that the IBM based speech separation elevates speech intelligibility in noisy environment. In our idea, the framework of supervised speech separation method can summarized into three important components, that is, acoustic features extraction, machine learning approach and training targets. Different algorithms focus on above components. Rickard proposed the degenerate unmixing estimation technique (DUET) algorithm [Rickard (2007)] to classify TF units by interaural time difference (ITD) and interaural level difference (ILD) for target signal separation. Roman et al. [Roman, Wang and Brown (2004)] utilized two binaural features to estimate IBM based on maximum a posteriori probability (MAP). Alinaghi et al. [Alinaghi, Wang and Jackson (2011)] combines binaural clues with blind source separation algorithms for speech separation in reverberant environment. The interaural phase difference (IPD) and ILD are modeled by a Gaussian mixture model (GMM), which is used to evaluate the classification of each TF unit. Abdipour et al. [Abdipour, Akbari, Rahmani et al. (2015)] proposed a speech separation algorithm based on spatial cues and model adaptation. They follow a maximum likelihood linear regression (MLLR) approach for tracking source relocations. Recently, the deep neural network (DNN) based method has made significant progress in speech separation. Wang et al. [Wang and Wang (2013)] proposed the DNN-SVM system to deal with speech separation under reverberant and noisy conditions. DNN is trained to extract more discriminant features, and then SVM is for sub-band IBM estimation. DNN-SVM system significantly improved speech intelligibility. Wang firstly used DNN to binaural speech separation [Jiang and Wang (2014)]. In each TF unit, the spatial feature ITD and ILD, and the monaural feature GFCC are extracted as input features to train the DNN. This study shows that a trained DNN generalizes well to the untrained spatial configurations of sound sources, that is, the specific placement of sound sources and sensors in an acoustic environment. Also, when the target and the interference source are co-located or close to each other, the monaural features improve separation performance. The spectral monaural feature [Zhang and Wang (2017)] is extended to complementary monaural feature set including amplitude modulation spectrum (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP) and mel-frequency cepstral coefficients (MFCC). Also DNN training aims to estimate the ideal ratio mask (IRM). Xiao et al. [Xiao, Zhao, Nguyen et al. (2016)] used DNN to predict static parameters, differential parameters and cross-correlation features, which improves the speech separation performance for reverberant and noisy speech. High-level features [Yu, Wang and Han (2016)] are extracted from low-level features, such as mixing vector (MV), ILD and IPD by unsupervised learning, and then supervised learning are to find the nonlinear functions between high-level features and the orientations of dominant source. Based on trained networks, the probability that each TF unit belongs to different sources (target and interferers) can be estimated based on the localization cues which is further used to generate the soft mask for source separation. Two-stage DNN structure is proposed in Zhao et al. [Zhao, Wang and Wang (2017)]. The masking from the first DNN is used for noise reduction, and the second DNN is spectral mapping for dereverberation. The results show that the performance of the two-stage DNN is greatly improved compared to the single-stage DNN. Bi-directional long short term memory (BLSTM) [Wang, Zhang and Wang (2018, 2018)] is also utilized to determine whether or not the TF unit is dominated by target speech. Then TF units containing clean phase is for DOA estimation. Also, the classification of TF unit [Wang and Wang (2018)] is determined by deep clustering and permutation, integrates spectral and inter-channel phase patterns for multichannel speech separation.

We note that, in Wang et al. [Wang and Chen (2018); Chen and Wang (2017); Ding, Li, Han et al. (2019)], LSTM shows its powerful ability to capture long-term speech contexts for speaker and noise-independent speech enhancement. Inspired by these researches, our study is conducted to use LSTM for binaural speaker separation, where LSTM framework is designed combining with spatial features and modified IRM. Hereafter, we treat the speech separation problem as a speaker separation problem, since we pursue the speech of target speaker by using the spatial information of speaker. In our scheme, since spatial features of consecutive frames are related for un-moving or slow-moving speakers, binaural spatial features are trained by BLSTM. In detail, Cross-Correlation Functions (CCF) with ITD and ILD are calculated at TF unit level as spatial features. Then, BLSTM classifier is trained at each frequency channel for the frequency-varied binaural features. Moreover, assuming the sum of each speaker magnitude is consistent with the original mixture, we force the BLSTM outputs are the proportions of the target speech for their corresponding mixtures, where the training labels are provided by the modified IRM.

The remainder of the paper is organized as follows. Section 2 presents an overview of our

BLSTM-based binaural speech separation system and extraction of spatial features. Section 3 describes the structure and training of BLSTM networks. The simulation results and analysis are provided in Section 4. The conclusion is drawn in Section 5.

2 Structure system overview and feature extraction

The proposed speaker separation system is illustrated in Fig. 1. The binaural signals are first decomposed into TF units independently by 33-channel Gammatone filters. CCF, IID and ILD are extracted in each TF units, and regarded as spatial features. The BLSTM is trained to estimation IRM by these spatial features. Target speech is reconstructed from IRM and the mixture.

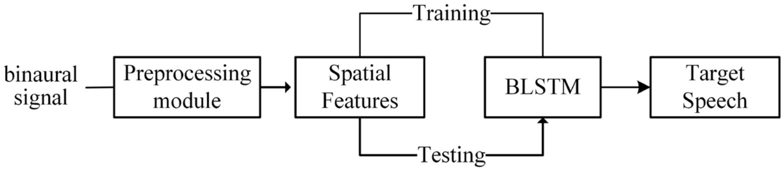

Figure 1: Schematic diagram of the BLSTM based binaural separation system

The physical model for binaural speech signals in reverberant and noisy environments can be formulated as:

$$
\begin{aligned}
x_L(t) &= h_{1,L} * s_1(t) + h_{2,L} * s_2(t) + n_L(t) \\
x_R(t) &= h_{1,R} * s_2(t) + h_{2,R} * s_2(t) + n_R(t)
\end{aligned}
\qquad (1)
$$

where $x_L(t)$ and $x_R(t)$ are defined as binaural mixture signals, $s_1(t)$ and $s_2(t)$ represent two speech sources, h_L and h_R are the Binaural Room Impulse Response (BRIR) for left and right ears respectively for each speech source; Moreover, symbols $n_L(t)$ and $n_R(t)$ are additive noise for each ear, which are irrelevant to each other.

Both left-ear and right-ear signal, $x_L(t)$ and $x_R(t)$, are decomposed into cochleagrams. The central frequencies of Gammatone filters ranges from 50 Hz to 8000 Hz on the equivalent rectangular bandwidth (ERB). The output of each channel is divided into 32-ms frame length with 16-ms frame shift. The binaural signals are then converted into TF units. In each unit, CCF, IID and ITD between the left-ear and right-ear signals are exacted.

The normalized CCF of a TF unit pair is defined as:

$$
CCF(i,k,d) = \frac{\sum_{m=0}^{N-1} x_L(i,k,m) x_R(i,k,m+d)}{\sqrt{\left[\sum_{m=0}^{N-1} \left(x_L(i,k,m)\right)^2\right]\left[\sum_{m=0}^{N-1} \left(x_R(i,k,m)\right)^2\right]}} \quad -L \le d \le L \qquad (2)
$$

where $x_L(i,k,m)$ and $x_R(i,k,m)$ are the binaural signals of TF unit at *i-th* channel and *k-th* frame, m is the sample number in a TF unit; N is the frame length; d denotes the delay between binaural signals and of the range from [-1 1] ms. For the 16 kHz sampling rate, the value of L is set to 16 with the dimension of CCF of size 33.

The ITD of each TF unit is the delay corresponding to the maximum value of CCF. It is formulated as:

$$ITD(i,k) = \arg\max_{d} CCF(i,k,d), \quad -L \le d \le L \tag{3}$$

And the ILD is defined as the energy ratio of the left and right ears in each TF unit pair:

$$ILD(i,k) = 20\log_{10} \frac{\sum_{m}\left(x_R\left(i,k,m\right)\right)^2}{\sum_{m}\left(x_L\left(i,k,m\right)\right)^2} \tag{2}$$

The spatial feature vector extracted in each TF unit pair is as follows:

$$F(i,k) = [CCF(i,k,-L), CCF(i,k,-L+1),..., CCF(i,k,L), ITD(i,k), ILD(i,k)] \tag{3}$$

As the main spatial feature, the CCF for two sources with different azimuths is described in Fig. 2. With Head related impulse response (HRIR) and TIMIT data, those two source are located at -30° and 60° respectively. The upper half of the figure is a curve of CCF at each TF unit, while the lower half is a CCF curve of the all channels.

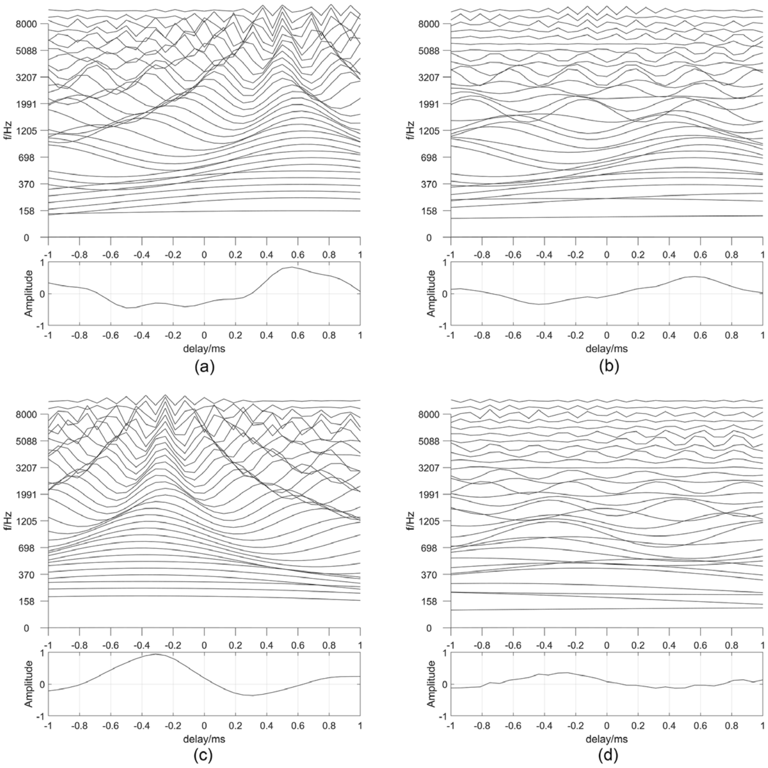

Figure 2: CCF of speech source with different azimuth in clean and noisy environment

Fig. 2 shows the CCFs in various acoustic environment. CCFs for speech source with azimuth 60° in an anechoic room and reverberant room are shown in Figs. 2(a) and 2(b) respectively. Figs. 2(c) and 2(d) show the CCF for -30° source in different acoustic environments. In both circumstances, the CCFs all have a similar peak in different TF units, which corresponds to the source location. In the high frequency channel, CCF has several peaks due to the phase wrapping. From Figs. 2(b) and 2(d), owing to the noise and reverberation, the peak of CCF is not obvious with the azimuth. Specifically, CCFs of each TF unit in reverberant room does not have obvious discriminability.

For unmoving or slow-moving speaker, since the spatial feature of consecutive frames are highly correlated, the LSTM can be used to model temporal dynamics of spatial features. Time-step is the important parameter of LSTM, which is related to the inter-frame correlation of the spatial features. Fig. 3 shows the inter-frame correlation coefficient of spatial features in different acoustic environments.

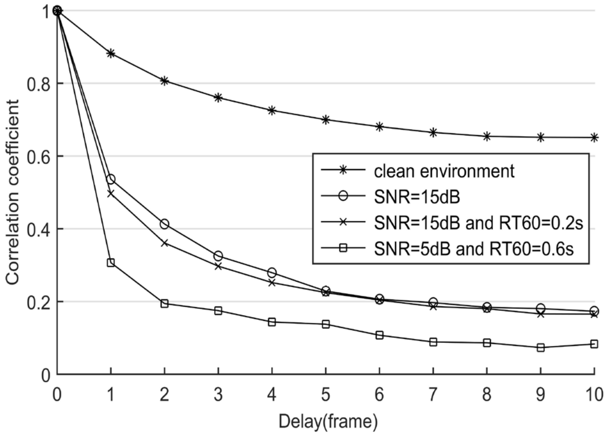

Figure 3: The inter-frame correlation coefficient of spatial features

In Fig. 3, the abscissa indicates the frame interval of the spatial feature, and the ordinate is the inter-frame correlation coefficient of the spatial feature. The selected environment includes a clean environment, a noisy environment (SNR=15 dB), a noisy and reverberant environment (SNR are set to 15 dB, 5 dB, reverberation times are 0.2 s and 0.6 s). As a result, spatial features show significant inter-frame correlation, whether in a clean environment or in a reverberant and noisy environment. Noise and reverberation reduce the inter-frame correlation. When the frame interval exceeds 6, the correlation coefficient of the spatial features will be less than 0.1. At this time, inter-frame correlation is too small to be ignored. Thus the time step in the LSTM is set to 11, That is, IRM is estimated by using the spatial features of consecutive 11 frames (5 before and 5 after the current TF unit).

3 BLSTM based speaker separation

3.1 Training targets

The IRM is defined as Wang et al. [Wang, Narayanan and Wang (2014)]：

$$IRM = \left(\frac{S(i,k)^2}{S(i,k)^2 + N(i,k)^2} \right)^{\beta} \tag{4}$$

where $S(i,k)^2$ and $N(i,k)^2$ denote speech energy and noise energy within a TF unit, respectively. The tunable parameter β is commonly set to 0.5.

In this paper, we separate the different speakers through spatial information, the IRM is defined as:

$$IRM = \frac{S_1(i,k)^2}{(S_1(i,k)+S_2(i,k)+N(i,k))^2} \tag{5}$$

where $S_1(i,k)$ and $S_2(i,k)$ represent the two speaker's signals, and $N(i,k)$ is the additive noise.

In Eq. (7), numerator represents the energy of the target speech, while the denominator is the total energy of mixture.

In a given TF unit, two sources and noise are regarded as irrelevant, IRM for different speaker and noise are rewritten as:

$$\begin{aligned} IRM_1 &= \frac{S_1(i,k)^2}{S_1(i,k)^2+S_2(i,k)^2+N(i,k)^2} \\ IRM_2 &= \frac{S_2(i,k)^2}{S_1(i,k)^2+S_2(i,k)^2+N(i,k)^2} \\ IRM_{noise} &= \frac{N(i,k)^2}{S_1(i,k)^2+S_2(i,k)^2+N(i,k)^2} \end{aligned} \tag{6}$$

The LSTM outputs the IRM which indicates the magnitude ratio of the sound source to the mixture, and Eq. (8) guarantees that the IRM sum of all LSTM output neurons for each channel is 1.

3.2 Speech separation and reconstruction

As for speakers with different direction, LSTM outputs the IRM corresponding to the each speaker in given azimuth. In binaural speaker separation, sound sources are only located in the front half of the horizontal plane. With the MIT HRIR, the azimuth is uniformly sampled with the steps of 10°, the front plane has 19 directions, corresponding to the 19 output neurons of the LSTM network. Also, for the ambient noise, LSTM is designed with an additional noise term corresponding to the IRM of the noise in the mixture. Therefore, the number of LSTM output neurons is 20. Thus, the training target of LSTM for each channel is the IRM vector, including Eq. (8), that is:

$$IRM_{vector} = [0,...,IRM_1,...,0,...,IRM_2,...,IRM_{noise}] \tag{9}$$

where dimension of IRM_{vector} is 20×1, IRM_1, IRM_2 and IRM_{noise} are the IRM for two speaker sources and noise. The positions of IRM_1 and IRM_2 in the vector are consistent with the azimuth of the speech source.

According to the IRM of LSTM output neuron, the target speech in each TF unit is reconstructed by:

$$\begin{aligned} \tilde{s}_1(i,k) &= IRM_1(i,k) \times x(i,k) \\ \tilde{s}_2(i,k) &= IRM_2(i,k) \times x(i,k) \end{aligned} \tag{10}$$

where IRM_1(i,k)and IRM_2(i,k) are the estimated *IRM* from LSTM for two speakers.

3.3 The architecture of BLSTM

The LSTM network consists of an input layer, an LSTM network layer and an output layer. The LSTM network layer is composed of a bidirectional long and short time memory unit (BiLSTM). Each memory unit is bidirectionally connected with the front and latter memory unit. The dimension of the input layer is equal to the dimension of spatial feature. The time-step in this paper is set to 11 (5 before and 5 after the current TF unit). The BLSTM network of the proposed system has two layers, including the hidden layer of 256 neurons and the output layer of 20 neurons. The structure of the LSTM is shown in Fig. 4.

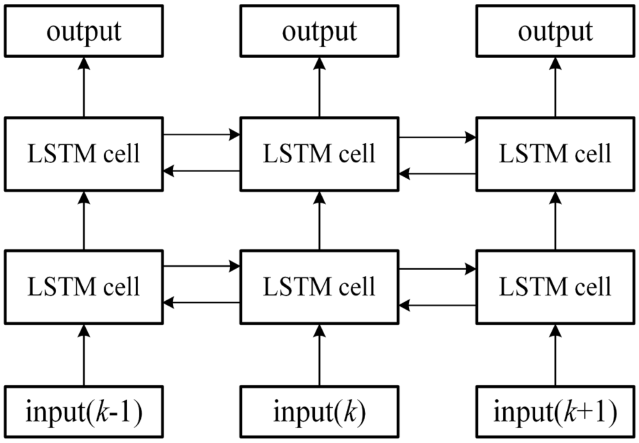

Figure 4: The structure of BLSTM

Initialize the weight matrix in each LSTM network memory unit, the IRM vector in Eq. (9) is employed as the training target. The cost function here is the mean square error (MSE) between IRM vector and output of LSTM, which is calculated as follows:

$$J = \frac{1}{2} E\left[\left\| IRM'_{vector} - IRM_{vector} \right\|_2^2 \right] \tag{7}$$

where E [·] denotes the expectation operation, $\|\cdot\|_2$ represents $L2$ norm; IRM'_{vector} is the output of LSTM, IRM_{vector} is the desired output of LSRM.

In this paper, the total number of training epochs is 20, the learning rate is 0.003. Adam optimizer is used to optimize the learning rate.

4 Simulation and result analysis

4.1 Simulation setup

For both training and test, the mono source signals taken from the CHAINS Speech Corpus [Cummins, Grimaldi, Leonard et al. (2006)], is convoluted with the MIT HRIR to generate binaural signals. The CHAINS speech corpus contains 33 sentences spoken by 36 speakers. 9 sentences are selected from the CSLU Speak Identification corpus and 24 sentences are from the TIMIT corpus. Binaural signals of two source with different azimuth are mixed to generate the mixture speech. One of the source is male speech and the other is a female speech. The speakers and speech content for training differs from that for test.

Also, the Gaussian white noise is added to the binaural mixture speech as the ambient noise. The noise is uncorrelated with the binaural signals. In addition, binaural noise is uncorrelated with each other. The SNR of the mixtures signals for training and test is set to 0, 5, 10, 15 and 20 dB. For SNR generalization, the SNR for test also includes -3, 3, 6, 9 and 12 dB.

Binaural room impulse response (BRIR) is obtained by ROOMSIM software [Campbell, Palomaki and Brown (2005)], which simulate room acoustics. The reverberation time (RT60) of BRIR is 0.2 s and 0.6 s. The reverberation signals are only used for test, which verifies the generalization of the proposed algorithm to reverberation.

The two speech sources are located on the front half of the horizontal plane with different azimuth. There are 171 combinations of source spatial configuration, all of which are used for training and test. The placement of speech sources and receiver are depicted in Fig. 5.

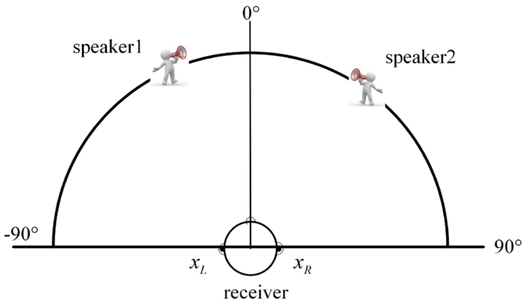

Figure 5: The spatial configuration of sources and receiver

Perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) are used to evaluate the performance of speaker separation. STOI is intelligibility metric, the value range is typically between 0 and 1. PESQ is correlated with the speech quality, the value is in a range of -0.5 to 4.5.

We compare the performance of the proposed method, LSTM based separation with IRM, with several other related methods for binaural speech separation. The three algorithms involved in the comparison are: the speech separation algorithm based on DUET, DNN based method with IBM, and DNN based method with traditional IRM.

4.2 Evaluation and analysis

Firstly, we evaluate the performance of the proposed algorithm in the matched noisy environment, that is, the test mixtures and the training mixtures have the same SNR. PESQ for different algorithm are shown in Tab. 1.

Table 1: PESQ of different methods in noisy environment

SNR (dB)	DUET	IBM-DNN	IRM-DNN	IRM-LSTM
0	1.403	1.467	1.946	1.874
5	1.57	1.656	2.121	2.140
10	1.754	1.834	2.258	2.355
15	1.923	1.982	2.386	2.528
20	2.102	2.119	2.510	2.654
clean	2.628	2.355	2.765	2.795

According to Tab. 1, the proposed method outperforms the other comparison methods in noisy environment. The second best system is IRM-DNN. The IRM-LSTM and IRM-DNN is much better than IBM-DNN, indicates that soft label is more suitable for the training target for speaker separation. The proposed method takes the correlation of spatial features of consecutive TF units, which gets the further performance gains.

As the two best algorithms for speech separation, only IRM-LSTM and IRM-DNN are compared in the following. The STOI is shown in Tab. 2.

Table 2: STOI of IRM-DNN and IRM-LSTM in noisy environment

SNR (dB)	IRM-DNN	IRM-LSTM
0	0.574	0.603
5	0.673	0.684
10	0.720	0.735
15	0.741	0.765
20	0.752	0.782
clean	0.762	0.796

From Tab. 2, the proposed algorithm also obtains the STOI gains in all SNR conditions.

The above results are the performance comparison under the condition that the test environment is matched with training environment. Below we give the PESQ results in an unmatched SNR environment, which SNR of test environment is set to -3, 3, 6, 9 and 12 dB. The result is shown in Tab. 3.

Table 3: PESQ comparison in the unmatched SNR

SNR (dB)	IRM-DNN	IRM-LSTM
-3	1.939	1.867
3	2.143	2.161
6	2.221	2.322
9	2.293	2.452
12	2.371	2.552

From Tab. 3, we can see that in unmatched SNR environments, the proposed system achieves the best results. Although the test SNRs aren’t included in the domain of training situation, IRM-LSTM still has reliable separation performance, indicating that the proposed algorithm is generalized to unmatched SNR environment.

In addition to ambient noise, reverberation is also a common interference signal. The reverberant signals are not used during the training. But in the test, we used BRIR to generate the reverberated mixtures to verify algorithm generalization to reverberation. RT60s take the values of 0.2 s and 0.6 s. The results of STOI and PESQ are shown in Tab. 4, 5, 6 and 7, respectively.

Table 4: PESQ comparison under reverberant environment (RT60 is 0.2 s)

SNR (dB)	IRM-DNN	IRM-LSTM
0	1.717	1.710
5	1.971	2.004
10	2.139	2.151
15	2.262	2.359
20	2.345	2.380

Table 5: STOI comparison under reverberant environment (RT60 is 0.2 s)

SNR (dB)	IRM-DNN	IRM-LSTM
0	0.545	0.601
5	0.646	0.669
10	0.711	0.717
15	0.745	0.747
20	0.763	0.760

Table 6: PESQ comparison under reverberant environment (RT60 is 0.6 s)

SNR (dB)	IRM-DNN	IRM-LSTM
0	1.664	1.645
5	1.913	2.024
10	2.069	2.120
15	2.180	2.253
20	2.252	2.298

Table 7: STOI comparison under reverberant environment (RT60 is 0.6 s)

SNR (dB)	IRM-DNN	IRM-LSTM
0	0.526	0.582
5	0.621	0.650
10	0.686	0.692
15	0.721	0.719
20	0.739	0.732

Based on the results in Tabs. 4-7, we found that both methods achieved reliable speech quality and intelligibility under reverberant conditions. Compared to Tabs. 1 and 2 with no reverberation, PESQ and STOI are slightly reduced, but the performance is still stable. At the same time, the performance of IRM-LSTM is better than that of IRM-DNN, which means that IRM-LSTM has better generalization performance to reverberation.

5 Conclusion

In this work, we present a LSTM-based binaural speech separation framework. By considering the temporal correlation of spatial features, we estimate the IRM for each sound source in TF unit more accurately by LSTM model. The LSTM-based speech separation has shown its ability to improve speech quality and intelligibility. Also, the proposed algorithm shows consistent results in unmatched reverberant and noisy conditions. The generalization ability is due to the use of LSTM model.

Acknowledgment: This work is supported by the National Nature Science Foundation of China (NSFC) under Grant Nos. 61571106, 61501169, 41706103 and the Fundamental Research Funds for the Central Universities under Grant No. 2242013K30010.

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

References

Abdipour, R.; Akbari, A.; Rahmani, M.; Nasersharif, B. (2015): Binaural source separation based on spatial cues and maximum likelihood model adaptation. *Digital Signal Processing*, vol. 36, pp. 174-183.

Alinaghi, A.; Wang, W.; Jackson, P. J. (2011): Integrating binaural cues and blind source separation method for separating reverberant speech mixtures. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 209-212.

Benesty, J.; Chen, J.; Huang, Y. (2008): *Microphone Array Signal Processing*. Springer Science & Business Media.

Campbell, D.; Palomaki, K.; Brown, G. (2005): A matlab simulation of "shoebox" room acoustics for use in research and teaching. *Computing and Information Systems*, vol. 9, no. 4, pp. 48.

Chen, J.; Wang, D. (2017): Long short-term memory for speaker generalization in supervised speech separation. *Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705-4714.

Cummins, F.; Grimaldi, M.; Leonard, T.; Simko, J. (2006): The chains corpus: characterizing individual speakers. *Proceedings of SPECOM*, vol. 6, pp. 431-435.

DiBiase, J. H.; Silverman, H. F.; Brandstein, M. S. (2001): Microphone arrays: signal processing techniques and applications. *Chapter Robust Localization in Reverberant Rooms*, pp. 157-180.

Ding, L.; Li, L.; Han, J.; Fan, Y.; Hu, D. (2019): Detecting domain generation algorithms with Bi-LSTM. *Computers, Materials & Continua*, vol. 61, no. 3, pp. 1285-1303.

Han, K.; Wang, D. (2012): A classification based approach to speech segregation. *Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475-3483.

Jarrett, D. P.; Habets, E. A.; Naylor, P. A. (2017): *Theory and Applications of Spherical Microphone Array Processing*. Springer, USA.

Jiang, Y.; Wang, D.; Liu, R.; Feng, Z. (2014): 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.

Kumar, K.; Kim, C.; Stern, R. M. (2011): Delta-spectral cepstral coefficients for robust speech recognition. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4784-4787.

Rickard, S. (2007): The DUET blind source separation algorithm. *Blind Speech Separation*, pp. 217-241.

Roman, N.; Wang, D.; Brown, G. J. (2004): A classification-based cocktail-party processor. *Advances in Neural Information Processing Systems*, pp. 1425-1432.

Sawada, H.; Araki, S., Mukai, R.; Makino, S. (2006): 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.

Shao, Y.; Wang, D. (2008): Robust speaker identification using auditory features and computational auditory scene analysis. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1589-1592.

Sinex, D. G. (2013): Recognition of speech in noise after application of time-frequency masks: dependence on frequency and threshold parameters. *Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2390-2396.

Su, J.; Tao, H.; Tao, M.; Wang, D.; Xie, J. (2017): 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.

Wang, D. (2008): Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in Amplification*, vol. 12, no. 4, pp. 332-353.

Wang, D.; Brown, G. J.; Darwin, C. (2008): Computational auditory scene analysis: principles, algorithms and applications. *Acoustical Society of America Journal*, vol. 124, no. 1, pp. 13.

Wang, D.; Chen, J. (2018): 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.

Wang, Y.; Narayanan, A.; Wang, D. (2014): On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849-1858.

Wang, Y.; Wang, D. (2013): Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381-1390.

Wang, Z. Q.; Wang, D. (2018): 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.

Wang, Z. Q.; Zhang, X.; Wang, D. (2018): Robust speaker localization guided by deep learning-based time-frequency masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178-188.

Wang, Z. Q.; Zhang, X.; Wang, D. (2018): Robust TDOA estimation based on time-frequency masking and deep neural networks. *Proceedings of INTERSPEECH*, pp. 322-326.

Xiao, X.; Zhao, S.; Nguyen, D. H. H.; Zhong, X.; Jones, D. L. et al. (2016): Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation. *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 4, pp. 1-18.

Yu, Y.; Wang, W.; Han, P. (2016): 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.

Zhang, X.; Wang, D. (2017): 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.

Zhao, Y.; Wang, Z. Q.; Wang, D. (2017): A two-stage algorithm for noisy and reverberant speech enhancement. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5580-5584.