INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Deep Recurrent Neural Network based Monaural Speech Separation using Recurrent Temporal Restricted Boltzmann Machines

Suman Samui, Indrajit Chakrabarti, Soumya K Ghosh

Indian Institute of Technology Kharagpur, India
samuisuman@gmail.com, indrajit@ece.iitkgp.ernet.in, skg@cse.iitkgp.ernet.in

Abstract

This paper presents a single-channel speech separation method implemented with a deep recurrent neural network (DRNN) using recurrent temporal restricted Boltzmann machines (RTRBM). Although deep neural network (DNN) based speech separation (denoising task) methods perform quite well compared to the conventional statistical model based speech enhancement techniques, in DNN-based methods, the temporal correlations across speech frames are often ignored, resulting in loss of spectral detail in the reconstructed output speech. In order to alleviate this issue, one RTRBM is employed for modelling the acoustic features of input (mixture) signal and two RTRBMs are trained for the two training targets (source signals). Each RTRBM attempts to model the abstractions present in the training data at each time step as well as the temporal dependencies in the training data. The entire network (consisting of three RTRBMs and one recurrent neural network) can be fine-tuned by the joint optimization of the DRNN with an extra masking layer which enforces a reconstruction constraint. The proposed method has been evaluated on the IEEE corpus and TIMIT dataset for speech denoising task. Experimental results have established that the proposed approach outperforms NMF and conventional DNN and DRNN-based speech enhancement methods.

Index Terms: Speech separation, deep recurrent neural network, recurrent temporal restricted Boltzmann machines, deep learning.

1. Introduction

Monaural speech separation is important for a wide range of applications, such as speech communication, speech denoising task, singing-voice separation, digital hearing aid etc. In this paper, we mainly focus on the problem of separating the clean speech from back-ground noise or interfering speech from a single-channel microphone recording. Recently, the data-driven supervised learning based approach [1] of speech separation has demonstrated remarkable improvement compared to the conventional speech enhancement methods [2][3][4][5][6][7][8].

Based on the type of training targets, these supervised training based techniques can be categorized into two major classes: (i) $masking\text{-}based$ methods [9] where a classifier is trained to learn a mapping function from mixed signal to a time-frequency (T-F) mask such as IBM (Ideal Binary Mask) or IRM (Ideal Ratio Mask) and the estimated mask is used to separate out the clean speech from the mixture. Due to the recent success of deep learning, it has been observed that DNN-based techniques [10][11][12] have outperformed earlier the GMM and SVM based speech separation methods [13][14].

On the other hand, (ii) spectral $mapping\text{-}based$ methods train DNN as a non-linear regression model to perform speech separation task [15][16]. Multiple-target based approaches [17][18] had also been proposed in which the DNN model is used to estimate not only the target speech but also the interfering speech (or noise). It showed that using dual outputs, the perceived quality of speech separation could be improved. Huang $etal$. [19] also proposed a speech separation technique which jointly models all the sources within a mixture as targets to a deep recurrent neural network (DRNN).

In all these DNN-based methods, the temporal correlations across speech frames are often ignored, resulting in spectral detail loss in the reconstructed output speech in the time domain. As speech signal is a highly structured signal, leveraging temporal context is always an important criterion for improving the performance of any speech processing task. Generally, in conventional DNN-based speech separation techniques, instead of using a single frame, the concatenation of neighbouring frames using different sizes of contextual windows is often used in a learning machine as its input for predicting the output [10][19]. In the current work, in order to leverage temporal information in speech, a single-channel speech separation framework using recurrent temporal restricted Boltzmann machines (RTRBM) has been proposed to jointly model all the sources within a mixture as targets to a deep recurrent neural network (DRNN). A RTRBM [20] is a non-linear probabilistic model which can receive a collection of time-series feature vectors with its memory models, allowing it to capture the temporal information in the high-order feature space where speech features are converted more easily than in an original acoustic feature space. In the method, one RTRBM is employed for modelling the acoustic features of input (mixture) signal and two RTRBMs are trained for the two training targets (source signals) i.e. speech signal and background noise. Each RTRBM attempts to model the abstractions present in the training data at each time step as well as the temporal dependencies in the training data. The entire network (consisting of three RTRBMs and recurrent neural network) can be fine-tuned by the joint optimization of the DRNN with an extra masking layer which enforces a reconstruction constraint. Experimental results show that the proposed approach outperforms NMF (Non-negative matrix factorization) based and conventional DNN and DRNN-based speech enhancement methods.

The rest of the article is organized as follows. In Section 2, we briefly recapitulate the fundamental theory of RTRBM. The proposed DRNN-based speech separation system using RTRBM is presented in Section 3. In Section 4, the experimental results are described and finally, Section 5 concludes the work.

2. Probabilistic models: RBM and RTRBM

RBM is a generative model which is represented by an undirected stochastic graphical model (also known as Markov Random Field), consisting of a pair of set of visible and hid-

http://dx.doi.org/10.21437/Interspeech.2017-57

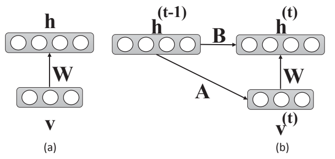

Figure 1: *Graphical representation of (a) an RBM and (b) an RTRBM.*

den units. RBM is initially introduced for unsupervised pretraining of deep belief network [21]. RTRBM [20] is an extended version of RBM model which can capture the temporal dependencies of sequential data by making hidden units receive additional input from the previous states of hidden units dynamically as shown in Figure 1. To represent and model the complicated distribution of the real-valued acoustic features of speech signals, we have considered the Gaussian-Bernoulli RTRBM (GB-RTRBM) where the visible units are continuous and follow a normal distribution and the hidden units follow a binary distribution. Given a current state of hidden units: $\mathbf{h}^{(t)} = [h_1^{(t)}, h_2^{(t)}, ..., h_J^{(t)}]$, $h^{(t)} \in \{0, 1\}$, a previous state of hidden units: $\mathbf{h}^{(t-1)} = [h_1^{(t-1)}, h_2^{(t-1)}, ..., h_J^{(t-1)}]$, $h^{(t-1)} \in \{0, 1\}$ and a current visible vector: $\mathbf{v}^{(t)} = [v_1^{(t)}, v_2^{(t)}, ..., v_I^{(t)}]$, $v^{(t)} \in \{0, \infty\}$ at the current frame t, the conditional probability can be defined with the help of an energy function $E(\mathbf{v}^{(t)}, \mathbf{h}^{(t)}|\mathbf{h}^{(t-1)})$ as follows:

$$p(\mathbf{v}^{(t)}|\mathbf{h}^{(t-1)}) = \frac{1}{Z}\sum_{\mathbf{h}^{(t)}} e^{-E(\mathbf{v}^{(t)},\mathbf{h}^{(t)}|\mathbf{h}^{(t-1)})} \tag{1}$$

$$E(\mathbf{v}^{(t)}, \mathbf{h}^{(t)}|\mathbf{h}^{(t-1)}) = -\mathbf{b}^{\mathbf{T}}(\frac{\mathbf{v}^{(t)}}{\sigma^{\mathbf{2}}}) - \mathbf{c}^{\mathbf{T}}\mathbf{h}^{(t)} - (\frac{\mathbf{v}^{(t)}}{\sigma^{\mathbf{2}}})^T\mathbf{W}\mathbf{h}^{(t)} - (\mathbf{h}^{(t-1)})^T\mathbf{A}(\frac{\mathbf{v}^{(t)}}{\sigma^{\mathbf{2}}}) - (\mathbf{h}^{(t-1)})^T\mathbf{B}\mathbf{h}^{(t)} \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{I\times J}$, $\mathbf{A} \in \mathbb{R}^{J\times I}$ and $\mathbf{B} \in \mathbb{R}^{J\times J}$ are the weight matrices between $\mathbf{v}^{(t)}$ and $\mathbf{h}^{(t)}$, $\mathbf{h}^{(t-1)}$ and $\mathbf{v}^{(t)}$, and $\mathbf{h}^{(t-1)}$ and $\mathbf{h}^{(t)}$, respectively. Moreover, $\mathbf{b} \in \mathbb{R}^{I\times 1}$ and $\mathbf{c} \in \mathbb{R}^{J\times 1}$ are the bias vectors of visible and hidden layers of RTRBM, respectively. $\sigma \in \mathbb{R}^{I\times 1}$ is the standard deviations associated with Gaussian visible units. The parameter $Z = \sum_{\mathbf{v}^{(t)}} \sum_{\mathbf{h}^{(t)}} exp(E(\mathbf{v}^{(t)}, \mathbf{h}^{(t)}|\mathbf{h}^{(t-1)}))$ is called the partition function (also known as normalization factor). I and J denote the number of units in the visible and hidden layers, respectively. The five parameters ($\mathbf{W}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{b}$, $\mathbf{c}$) of RTRBM are estimated by maximizing the log-likelihood function: $\mathcal{L} = \log \prod_t p(\mathbf{v}^{(t)}|\mathbf{h}^{(t-1)})$. Differentiating partially with respect to each parameter, the following equations can be obtained:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}_{ij}} =< \frac{v_i{}^{(t)}h_j{}^{(t)}}{\sigma_i^2} >_{data} - < \frac{v_i{}^{(t)}h_j{}^{(t)}}{\sigma_i^2} >_{model} \tag{3}$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{A}_{ij}} =< \frac{v_i{}^{(t)}h_j{}^{(t-1)}}{\sigma_i^2} >_{data} - < \frac{v_i{}^{(t)}h_j{}^{(t-1)}}{\sigma_i^2} >_{model} \tag{4}$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{B}_{j'j}} =< h_j{}^{(t)}h_{j'}{}^{(t-1)} >_{data} - < h_j{}^{(t)}h_{j'}{}^{(t-1)} >_{model} \tag{5}$$

$$\frac{\partial\mathcal{L}}{\partial b_i} =< \frac{v_i{}^{(t)}}{\sigma_i^2} >_{data} - < \frac{v_i{}^{(t)}}{\sigma_i^2} >_{model} \tag{6}$$

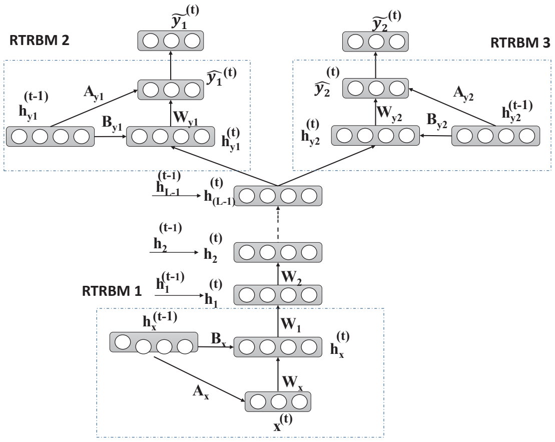

Figure 2: *Schematic diagram of the proposed DRNN-based speech separation system using RTRBM.*

$$\frac{\partial\mathcal{L}}{\partial c_j} =< h_j{}^{(t)} >_{data} - < h_j{}^{(t)} >_{model} \tag{7}$$

where $< . >_{data}$ and $< . >_{model}$ are the expectations of input data and inner model respectively. However, it is generally difficult to compute $< . >_{model}$. Contrastive divergence (CD) approximation of the gradient is used by replacing $< . >_{model}$ by running the Gibbs sampler. After estimating the aforementioned five parameters, the conditional probability of $\mathbf{h}^{(t)}$ given $\mathbf{v}^{(t)}$ and $\mathbf{h}^{(t-1)}$ and the conditional probability of $\mathbf{v}^{(t)}$ given $\mathbf{h}^{(t)}$ and $\mathbf{h}^{(t-1)}$ are given as:

$$p(h_j{}^{(t)} = 1|\mathbf{v}^{(t)}, \mathbf{h}^{(t-1)}) = \mathcal{S}(c_j^{(t)} + \mathbf{W}^{\mathbf{T}}_{\mathbf{:j}}(\frac{\mathbf{v}^{(t)}}{\sigma^{\mathbf{2}}}) + \mathbf{B}^{\mathbf{T}}_{\mathbf{:j}}\mathbf{h}^{(t-1)}) \tag{8}$$

$$p(v_i{}^{(t)} = v|\mathbf{h}^{(t)}, \mathbf{h}^{(t-1)}) = \mathcal{N}(v|b_i^{(t)} + \mathbf{W}^{\mathbf{T}}_{\mathbf{i:}}\mathbf{h}^{(t)}, \sigma_i^2) \tag{9}$$

where $\mathcal{S}(.)$ and $\mathcal{N}(.|\mu, \sigma^2)$ indicate an element-wise sigmoid function and Gaussian probability density function with the mean μ and variance σ^2 .

3. Proposed DRNN-based speech separation system

The method proposed in this article is a multiple-target based approach like in [18][19] where all the different sources (here we have assumed two sources) are separated out from the mixture signal by simultaneously modelling all the sources present in the mixture as the training targets using DRNN. In spite of directly using the acoustic features (spectral or log-mel filter-bank features) from a mixture to the DRNN, the input feature vector ($\mathbf{x^{(t)}}$) at frame t is projected into a latent feature space using RTRBM which captures the temporal dependencies in the features. Similarly, the feature space of training targets (sources $\mathbf{y_1^{(t)}}$ and $\mathbf{y_2^{(t)}}$) are transformed into a latent feature space using two RTRBMs. In our method, we have used three exclusive RTRBMS: (i) one RTRBM is adopted for the input mixture and (ii) the other two RTRBMs are used for the source signals in order to model the distribution of the acoustic features of input mixture ($\mathbf{x^{(t)}}$) and the two source signals ($\mathbf{y_1^{(t)}}$ and $\mathbf{y_2^{(t)}}$), respectively. These three RTRBMs are independently pre-trained in an unsupervised way as shown in Figure 2. At time t, the training input, $\mathbf{x}^{(t)}$ of the network is the acoustic features (spectral or log-mel filterbank features) from a mixture within an analysis window. The output predictions, $\mathbf{y_1^{(t)}}$ and $\mathbf{y_2^{(t)}}$ of the network are the spectral features of the two separated sources at time t. The hidden vectors: $\mathbf{h_x^{(t)}}$, $\mathbf{h_{y_1}^{(t)}}$, and $\mathbf{h_{y_2}^{(t)}}$ which can be considered as the high-order feature representation of the raw acoustic features, which are the hidden variables extracted from three RTRBMs, respectively. Therefore, the parameter set of our proposed method is given by: $\boldsymbol{\Theta} = \left\{\theta_{\mathbf{x}}, \theta_{\mathbf{y_1}}, \theta_{\mathbf{y_2}}, \theta_{\mathbf{n}}\right\}$, where $\theta_{\mathbf{x}} = \{\mathbf{W_x}, \mathbf{A_x}, \mathbf{B_x}, \mathbf{b_x}, \mathbf{c_x}\}$ are the parameters of the input RTRBM computed by using the training spectral features of the mixture signal, $\theta_{\mathbf{y_1}} = \{\mathbf{W_{y1}}, \mathbf{A_{y1}}, \mathbf{B_{y1}}, \mathbf{b_{y1}}, \mathbf{c_{y1}}\}$ and $\theta_{\mathbf{y_2}} = \{\mathbf{W_{y2}}, \mathbf{A_{y2}}, \mathbf{B_{y2}}, \mathbf{b_{y2}}, \mathbf{c_{y2}}\}$ are the parameters of the two output RTRBMs using the training spectral features of source signals, respectively. $\theta_{\mathbf{n}} = \{\mathbf{W_1}, \mathbf{W_2}, .., \mathbf{W_L}, \mathbf{d_1}, \mathbf{d_2}, ..., \mathbf{d_L}\}$ are the parameters of the DRNN trained using the extracted hidden variables, where $(L-1)$ is the number of hidden layers, d_l is the bias vector of l^{th} hidden layer, W_l is the weight matrix from the $(l-1)^{th}$ layer to the l^{th} layer. For the input mixture features, the parameters $\theta_{\mathbf{x}}$ is estimated by maximizing log-likelihood function $\mathcal{L} = \log \prod_t p(\mathbf{x}^{(t)} | \mathbf{h_x}^{(t-1)})$. After the parameters $\theta_{\mathbf{x}}$ are estimated, the latent features $\mathbf{h_x^{(t)}}$ are obtained using mean-field approximation as in (8). The same notion can be applied to the output features as well. The three projected vectors: $\mathbf{h_x^{(t)}}$, $\mathbf{h_{y_1}^{(t)}}$, and $\mathbf{h_{y_2}^{(t)}}$ can be given as follows

$$\mathbf{h_x^{(t)}} = \mathcal{S}(\mathbf{c_x^{(t)}} + \mathbf{W_x^T}(\frac{\mathbf{x}^{(t)}}{\sigma^{\mathbf{2}}}) + \mathbf{B_x^T h_x^{(t-1)}}) \tag{10}$$

$$\mathbf{h_{y_1}^{(t)}} = \mathcal{S}(\mathbf{c_{y_1}^{(t)}} + \mathbf{W_{y_1}^T}(\frac{\mathbf{y_1}^{(t)}}{\sigma^{\mathbf{2}}}) + \mathbf{B_{y_1}^T h_{y_1}^{(t-1)}}) \tag{11}$$

$$\mathbf{h_{y_2}^{(t)}} = \mathcal{S}(\mathbf{c_{y_2}^{(t)}} + \mathbf{W_{y_2}^T}(\frac{\mathbf{y_2}^{(t)}}{\sigma^{\mathbf{2}}}) + \mathbf{B_{y_2}^T h_{y_2}^{(t-1)}}) \tag{12}$$

where $\mathbf{h_x^{(0)}}$, $\mathbf{h_{y_1}^{(0)}}$ and $\mathbf{h_{y_2}^{(0)}}$ are assumed to be as zero vectors.

At the training phase of DRNN, the projected vectors of the spectral features of the mixture are the inputs of the neural network and the projected vectors of the spectral features of the two different sources are considered as the ground truth outputs. The output of the RNN is given by:

$$\mathbf{\tilde{h}_y^{(t)}} = f_o(\mathbf{z_l^{(t)}}) \tag{13}$$

where the following parameters can be defined as: $\mathbf{z_l^{(t)}} = \mathbf{W_l o_{(l-1)}^{(t)}} + \mathbf{U_l h_l}(\mathbf{z_l^{(t-1)}}) + \mathbf{d_l}$, $\mathbf{o_{(l-1)}^{(t)}} = \mathbf{f_h}(\mathbf{z_{(l-1)}^{(t)}})$ and $\mathbf{o_0^{(t)}} = \mathbf{h_x}(\mathbf{t})$. $\mathbf{\tilde{h}_y^{(t)}}$ is the concatenation of predicted projected vectors of two sources i.e. $\mathbf{\tilde{h}_{y1}^{(t)}}$ and $\mathbf{\tilde{h}_{y2}^{(t)}}$. $\mathbf{U_l}$ is the weight matrix for the recurrent connection at l^{th} layer. f_h and f_o are the activation function of hidden layers and output layer respectively. Finally, according to Eq.(11) and (12), using the mapping function of our method, the output of the neural network is given by:

$$\mathbf{\hat{y}_1^{(t)}} = \mathcal{S}(\mathbf{c_{y1}} + \mathbf{W_{y1}}\mathcal{S}(\mathbf{\tilde{h}_y^{(t)}}) + (\mathcal{S}(\mathbf{\tilde{h}_y^{(t-1)}}))^{\mathbf{T}} + \mathbf{A_{y1}}) \tag{14}$$

$$\mathbf{\hat{y}_2^{(t)}} = \mathcal{S}(\mathbf{c_{y2}} + \mathbf{W_{y2}}\mathcal{S}(\mathbf{\tilde{h}_y^{(t)}}) + (\mathcal{S}(\mathbf{\tilde{h}_y^{(t-1)}}))^{\mathbf{T}} + \mathbf{A_{y2}}) \tag{15}$$

It is useful to further smooth the source separation results with a time-frequency masking function which enforces the constraint that the sum of the prediction results is equal to the original mixture. The soft time-frequency mask can be defined as follows:

$$M_s^{(t)}(f) = \frac{\mathbf{\hat{y}_1^{(t)}}(f)}{\mathbf{\hat{y}_1^{(t)}}(f) + \mathbf{\hat{y}_2^{(t)}}(f)} \tag{16}$$

We can integrate the masking function into the neural network directly by adding an extra layer to the original output of the neural network as follows: $\tilde{y}_1^{(t)} = \dfrac{\mathbf{\hat{y}_1^{(t)}}}{\mathbf{\hat{y}_1^{(t)}} + \mathbf{\hat{y}_2^{(t)}}} \odot \mathbf{z^{(t)}}$ and $\tilde{y}_2^{(t)} = \dfrac{\mathbf{\hat{y}_2^{(t)}}}{\mathbf{\hat{y}_1^{(t)}} + \mathbf{\hat{y}_2^{(t)}}} \odot \mathbf{z^{(t)}}$ where $\mathbf{z^{(t)}}$ denotes the magnitude spectrum of the mixture signal and the operator $\odot$ indicates the element-wise multiplication (Hadamard product). In this way, we can integrate the constraints to the network and optimize the network with the masking function jointly. In order to secure a high signal to interference ratio (SIR), we have used a discriminative objective function [19] as follows:

$$||\mathbf{y_1^{(t)}} - \mathbf{\tilde{y}_1^{(t)}}||^2 + ||\mathbf{y_2^{(t)}} - \mathbf{\tilde{y}_2^{(t)}}||^2 - \gamma||\mathbf{y_1^{(t)}} - \mathbf{\tilde{y}_2^{(t)}}||^2 - \gamma||\mathbf{y_2^{(t)}} - \mathbf{\tilde{y}_1^{(t)}}||^2 \tag{17}$$

where γ is a constant chosen by the performance on the development set. The time domain signals are reconstructed based on the inverse short time Fourier transform (ISTFT) of the estimated spectra.

4. Experimental results

The proposed framework is employed for speech de-noising, where one source is the clean speech and the other source is the background noise. The goal of the task is to separate clean speech from noisy speech. After conducting the pre-training of three RTRBMs independently on the input side and output side of the DRNN, we model the projected feature space of mixed signal using a DRNN model with four hidden layers each with 256 hidden units. The rectified linear activation function is used for the hidden units of the RNN. We optimize our models by back-propagating the gradients with respect to the training objectives. The limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization algorithm is used to train the models from random initialization. Empirically, it was found that the log-mel filterbank features provide worse performance than magnitude spectra, in this study, we have explored the magnitude spectra. The spectral representation is extracted using a 1024-point STFT with 50 % overlap. The frame size and frame shift are set to 20 msec and 10 msec respectively for STFT analysis.

4.1. Experimental setup and dataset

To evaluate the performance, we have created the test-stimuli by contaminating 720 Harvard sentences from IEEE corpus [22] and 600 TIMIT clean sentences [23] by six types of noises: babble, factory, domestic, cafeteria, street and speech-shaped noises(SSN). These noise instances have been taken from DEMAND (Diverse Environments Multichannel Acoustic Noise Database) corpus[24] and NOIZEOUS database [25]. All the speech signals are down-sampled at 16 KHz. All these noise samples are non-stationary in nature except the SSN which is stationary and produced synthetically with the same long-term spectrum as the sentences in the IEEE corpus. In order to create the mixture, each sentence of these two databases is mixed with 6 different noise instances at six different SNR levels (-6, -3, 0, +3, +6 and +9 dB). The training set uses 520 IEEE sentences and 500 TIMIT sentences (10 utterances from 50 speakers) with randomly selected segments from the first two minutes of a noise, while the test set uses another 200 IEEE sentences and 100 TIMIT sentences (10 utterances from 10 new speakers) with randomly selected segments from the remaining part of a noise. Therefore, the test set has different sentences and dif-

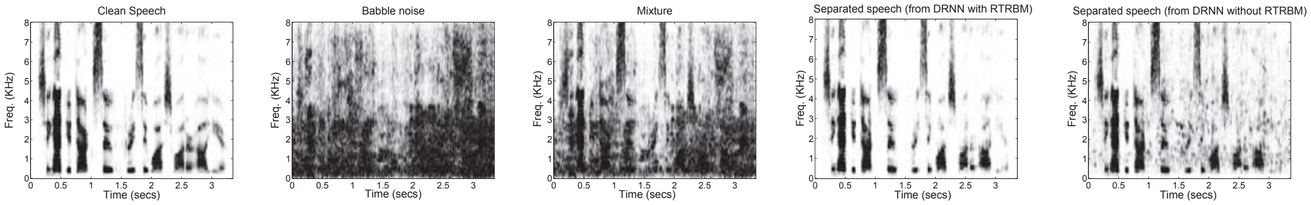

Figure 3: *A speech denoising example using magnitude spectrograms: (a) the clean speech for a test clip in IEEE corpus; (b) the babble noise; (c) mixture signal (speech + babble noise at -6 dB SNR; (d) the separated speech spectrogram from our proposed model (DRNN+RTRBM); (e) the separated speech spectrogram from DRNN without using RTRBM.*

ferent noise segments from the training set. An example of the separation results is shown in Figures 3.

4.2. Objective evaluation

In the current work, the speech separation (denoising task) has been quantitatively evaluated by BSS-EVAL metrics [26] which consists of three metrics: Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR), and Source to Distortion Ratio (SDR). Higher values of SDR, SAR, and SIR represent better separation quality. We have compared the proposed approach with the existing standard base-line speech separation approaches present in the literature such as (i) DRNN based method without RTRBMs [19], (ii) speech separation method using temporally regularized nonnegative matrix factorization (NMF) [27], (iii) DNN-based regression model [15] and (iv) DNN-SVM model [10]. The training target of the DNN-SVM model is set to IRM. A total of 7200 test mixtures for all SNR levels in the range of -6 dB to +9 dB, have been used for evaluating the performance of these systems in terms of BSS-EVAL metrics. The average comparative result is shown in Figure 4. It is worth noticing that the DRNN model with RTRBM pretraining performs quite well as compared to the aforementioned methods. We have also assessed the performance of the systems in terms of PESQ (perceptual evaluation of speech quality) and STOI (short-time objective intelligibility) [28] metrics which are highly correlated with the subjective listening test. Average PESQ and STOI scores have been shown in Table 1.

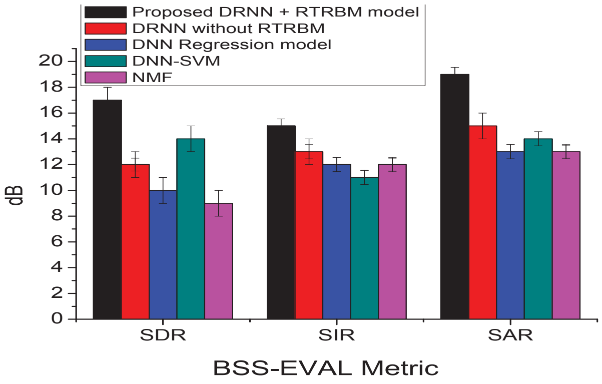

Figure 4: *Performance measure in terms of BSS-EVAL metrics.*

4.3. Subjective evaluation

The listening tests have been performed in a sound-proof room where the test stimuli and the enhanced speech files for each method were played to the listeners monoaurally through Sony headphone (Sony MDR-XB450AP) at their comfortable listening levels. A total of thirty sentences, evenly distributed between the male and female speakers, were selected for evaluation. The listeners were asked to rate each speech file on a 5-point scale. The experimental average Mean Opinion Score (MOS) is shown in Figure 5 for six SNR levels. As one may observe, there is a substantial improvement in intelligibility obtained with the proposed method, compared to that attained with unprocessed (noise-corrupted) speech and compared to the other methods.

Table 1: *Objective measure represented as PESQ(STOI) score*

SNR (dB)	Noisy Mixture	Proposed DRNN + RTRBM	DRNN without RTRBM	DNN-based Regression	NMF
-6	1.42(0.602)	**1.98(0.712)**	1.76(0.672)	1.64(0.656)	1.52(0.612)
-3	1.78(0.694)	**2.46(0.749)**	2.06(0.719)	2.06(0.692)	2.36(0.707)
0	2.06(0.762)	**2.78(0.816)**	2.38(0.765)	2.28(0.712)	2.46(0.752)
3	2.56(0.812)	**3.19(0.842)**	2.82(0.832)	2.66(0.749)	2.93(0.729)
6	3.16(0.842)	**3.52(0.882)**	3.36(0.868)	2.76(0.782)	3.14(0.712)
9	3.47(0.872)	**3.96(0.907)**	3.82(0.892)	3.26(0.832)	3.21(0.818)

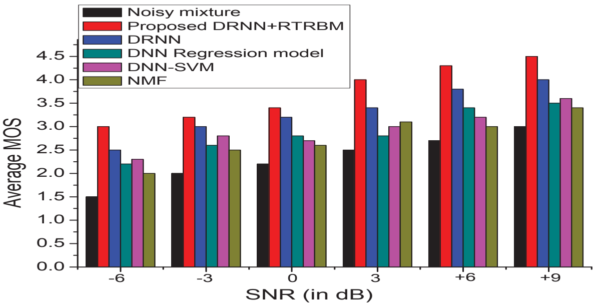

Figure 5: *Subjective listening test.*

5. Conclusions

In this article, we have proposed a speech separation system based on DRNN which utilizes RTRBMs to discover deep temporal correlations among speech frames and projects the input and output acoustic features of signals into a latent feature space. The entire network has been fine-tuned by the joint-optimization of the DRNN with an extra masking layer which enforces a reconstruction constraint. Experimental results have demonstrated that the proposed approach can outperform the other existing deep learning based speech separation methods.

6. References

[1] M. Kolbk, Z. H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, Jan 2017.

[2] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[5] R. C. Hendriks, T. Gerkmann, and J. Jensen, "Dft domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, 2013.

[6] S. Samui, I. Chakrabarti, and S. K. Ghosh, "Improved single channel phase-aware speech enhancement technique for low signal-to-noise ratio signal," *IET Signal Processing*, vol. 10, no. 6, pp. 641–650, 2016.

[7] P. Krishnamoorthy and S. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication*, vol. 53, no. 2, pp. 154–174, 2011.

[8] S. Samui, I. Chakrabarti, and S. K. Ghosh, "Two-stage temporal processing for single-channel speech enhancement," *Interspeech 2016*, pp. 3723–3727, 2016.

[9] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 1849–1858, 2014.

[10] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 1381–1390, 2013.

[11] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 2014, pp. 577–581.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 708–712.

[13] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[14] K. Han and D. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 1, pp. 7–19, 2015.

[16] S. Mirsamadi and I. Tashev, "Causal speech enhancement combining data-driven learning and suppression rule estimation," in *Interspeech 2016*, 2016, pp. 2870–2874.

[17] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Deep neural network based speech separation for robust speech recognition," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 532–536.

[18] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3734–3738.

[19] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[20] I. Sutskever, G. E. Hinton, and G. W. Taylor, "The recurrent temporal restricted boltzmann machine," in *Advances in Neural Information Processing Systems*, 2009, pp. 1601–1608.

[21] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[22] E. Rothauser, W. Chapman, N. Guttman, K. Nordby, H. Silbiger, G. Urbanek, and M. Weinstock, "Ieee recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, no. 3, pp. 225–246, 1969.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[24] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[25] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[27] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising." 2008.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2125–2136, 2011.