# DISCRIMINATIVE DEEP RECURRENT NEURAL NETWORKS FOR MONAURAL SPEECH SEPARATION

*Guan-Xiang Wang, Chung-Chien Hsu and Jen-Tzung Chien*

Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 30010, ROC

## ABSTRACT

Deep neural network is now a new trend towards solving different problems in speech processing. In this paper, we propose a discriminative deep recurrent neural network (DRNN) model for monaural speech separation. Our idea is to construct DRNN as a regression model to discover the deep structure and regularity for signal reconstruction from a mixture of two source spectra. To reinforce the discrimination capability between two separated spectra, we estimate DRNN separation parameters by minimizing an integrated objective function which consists of two measurements. One is the within-source reconstruction errors due to the individual source spectra while the other conveys the discrimination information which preserves the mutual difference between two source spectra during the supervised training procedure. This discrimination information acts as a kind of regularization so as to maintain between-source separation in monaural source separation. In the experiments, we demonstrate the effectiveness of the proposed method for speech separation compared with the other methods.

***Index Terms***— deep learning, discriminative learning, neural network, monaural speech separation

## 1. INTRODUCTION

Speech is one of the most important biosignals for human communication. Nowadays, many speech-related applications and devices have been developed to facilitate our daily lives. However, the system performance is usually deteriorated in adverse conditions. For example, it is important to conduct single-channel source separation to extract the target speech from a mixed noisy speech and use the enhanced speech in an automatic speech recognition system. The speech recognition performance is improved accordingly [1]. A typical instance of source separation problem is the cocktail-party problem [2, 3, 4] where the target speech is contaminated with a variety of interferences such as ambient noise, competing speech and background music [5]. Over the past few years, a number of single-channel separation algorithms have been proposed especially the model-based source separation approaches such as the non-negative matrix factorization (NMF) [6, 7, 8, 9, 10] and the deep neural network [11, 12, 13] based methods.

Recently, deep learning has emerged as a powerful machine learning approach. It produces state-of-the-art results in many research fields such as speech recognition and object detection. In general, deep learning adopts a hierarchical architecture to grasp latent information [14] from the given data for various classification and regression tasks. Deep neural network (DNN) can be developed and employed as a nonlinear approximation function to estimate the target speech signal from the mixed speech signal. For instance, DNN was applied to predict the separated spectra from the noisy spectra [15]. Extended from the standard DNN, deep recurrent neural network (DRNN) was proposed to explore the temporal information for source separation [13]. Later on, in [12], the long short-term memory (LSTM) was incorporated to tackle the gradient vanishing and exploding conditions in implementation of DRNN and exploit the long and short-term contextual information which helped the performance of monaural source separation.

In addition, the performance of model-based approaches can be improved with discriminative learning where the discrimination between source signals is optimized during the learning procedure. For example, discriminative NMF was proposed to adapt the NMF basis functions, which can optimize the separation tasks [9]. Another example is that different discriminative objectives of DRNN were explored to enhance the separation performance [13]. The discrimination information for two sources was taken into account as objective function for optimization in these two methods. In this paper, we present a new objective function to conduct the discriminative training (DT) for DRNN monaural source separation. The proposed new DT objective function consists of two measurements. One is the within-source reconstruction errors, which aims to separate two individual sources. The other term, which refers to the discrimination information, preserves the mutual difference between two individual sources. Such discrimination information regularizes the mismatched condition in the reconstructed spectra in two sources. The proposed criterion can improve the separation results in terms of source-to-distortion ratio (SDR) and source-to-interference ratio (SIR).

## 2. RELATED WORKS

### 2.1. Deep recurrent neural network

Deep neural network (DNN) is adopted as a nonlinear regression model to predict the magnitude spectra or the masking function of the separated signals given an input magnitude spectra of the mixed signal. A basic DNN model is composed of a chain of functional transformations.

Since the temporal dependency is known as an important information in time-series data such as audio signals, standard DNN does not take this information into account. The performance of source separation shall be constrained. Accordingly, the deep recurrent neural network (DRNN) is introduced to explore such dynamic temporal behavior. One basic calculation of DRNN is to feed the outputs of $l$-th layer from previous time instance $t-1$ into the current time step $t$ as

$$\mathbf{z}_t^{(l)} = f(\mathbf{a}_t^{(l)}) = f(\mathbf{w}^{(l)}\mathbf{z}_t^{(l-1)} + \mathbf{w}^{(ll)}\mathbf{z}_{t-1}^{(l)}) \quad (1)$$

where $\mathbf{a}^{(l)}$ and $\mathbf{z}^{(l)}$ denote the input and the output of an unit in $l$-th layer, respectively, and $f(\cdot)$ is the nonlinear activation function using sigmoid or ReLU [16]. $\mathbf{z}_t^{(l-1)}$ is the output of $(l-1)$-th layer at time $t$ and $\mathbf{z}_{t-1}^{(l)}$ is the output of $l$-th layer at time $t-1$, $\mathbf{w}^{(l)}$ denotes the weights between two layers at time $t$, and $\mathbf{w}^{(ll)}$ denotes the weights between two time steps, $t$ and $t-1$, of $l$-th hidden layer. The model parameters in DRNN consist of the weights in forward layers and recurrent layers $\mathbf{w} = \{\mathbf{w}^{(l)}, \mathbf{w}^{(ll)}\}$.

For the task of source separation, the fundamental objective function of DRNN model is formed as

$$E(\mathbf{w}) = \sum_{t=1}^{T}\left\{\frac{1}{2}\|\hat{\mathbf{x}}_{1,t}(\mathbf{w}) - \mathbf{x}_{1,t}\|^2 + \frac{1}{2}\|\hat{\mathbf{x}}_{2,t}(\mathbf{w}) - \mathbf{x}_{2,t}\|^2\right\} \quad (2)$$

where $\hat{\mathbf{x}}_{1,t}(\mathbf{w})$ and $\hat{\mathbf{x}}_{2,t}(\mathbf{w})$ denote the predicted spectra of two separated signals using DRNN parameters $\mathbf{w}$ and $\mathbf{x}_{1,t}$ and $\mathbf{x}_{2_t}$ denote the spectra of two true source signals, respectively. In this case, the objective function only considers the within-source reconstruction error in this regression model.

### 2.2. Discriminative learning

In order to regularize the reconstruction error, the discriminative measure was calculated and incorporated in the original objective function of Eq. (2) in a form of [13]

$$E(\mathbf{w}) = \sum_{t=1}^{T}\left\{\frac{1}{2}\|\hat{\mathbf{x}}_{1,t}(\mathbf{w}) - \mathbf{x}_{1,t}\|^2 + \frac{1}{2}\|\hat{\mathbf{x}}_{2,t}(\mathbf{w}) - \mathbf{x}_{2,t}\|^2 - \frac{\gamma}{2}\|\hat{\mathbf{x}}_{1,t}(\mathbf{w}) - \mathbf{x}_{2,t}\|^2 - \frac{\gamma}{2}\|\hat{\mathbf{x}}_{2,t}(\mathbf{w}) - \mathbf{x}_{1,t}\|^2\right\} \quad (3)$$

where $\|\cdot\|$ denotes the $\ell_2$ norm and $\gamma$ is the regularization parameter which adjusts the tradeoff between the regression/reconstruction error and the discrimination information. Here, the discrimination measurement in the last two terms of Eq. (3) is seen as the *between-source* information which is maximized to find the separated signals with the largest mutual information. This method was developed to increase the value of source-to-interference ratio (SIR) in the separated signals from monaural source separation [13].

## 3. DISCRIMINATIVE SOURCE SEPARATION

### 3.1. New objective function

In this paper, we present a new objective function for monaural source separation, which preserves the mutual difference between two source spectra during the separation procedure, based on

$$E(\mathbf{w}) = \sum_{t=1}^{T} E(\mathbf{w}_t) = \sum_{t=1}^{T}\Big\{\frac{1}{2}\|\hat{\mathbf{x}}_{1,t}(\mathbf{w}) - \mathbf{x}_{1,t}\|^2 + \frac{1}{2}\|\hat{\mathbf{x}}_{2,t}(\mathbf{w}) - \mathbf{x}_{2,t}\|^2 + \frac{\gamma}{2}\|\hat{\mathbf{d}}_t(\mathbf{w}) - \mathbf{d}_t\|^2\Big\}. \quad (4)$$

The discriminative measures using the *difference vectors* based on the reconstructed spectra $\hat{\mathbf{d}}_t(\mathbf{w})$ and the true spectra $\mathbf{d}_t$ are expressed by

$$\mathbf{d}_t = \mathbf{x}_{1,t} - \mathbf{x}_{2,t}, \quad \hat{\mathbf{d}}_t(\mathbf{w}) = \hat{\mathbf{x}}_{1,t}(\mathbf{w}) - \hat{\mathbf{x}}_{2,t}(\mathbf{w}) \quad (5)$$

where both information of angle and magnitude are considered. Basically, the first two terms in this discriminative objective function represent the within-source reconstruction errors due to individual source spectra $\{\mathbf{x}_1, \mathbf{x}_2\}$. The third term conveys the discrimination information which measures the mutual difference between two source spectra during the supervised training procedure. This discrimination information acts as a kind of regularization for between-source separation in monaural source separation.

### 3.2. Model learning procedure

Deep recurrent neural network (DRNN) is developed as the regression function to estimate the demixed signals $\{\hat{\mathbf{x}}_{1,t}(\mathbf{w}), \hat{\mathbf{x}}_{2,t}(\mathbf{w})\}$. The model architecture of DRNN with $L$ layers is shown in Figure 1 where a soft masking function is applied for monaural source separation. Model learning procedure with an error backpropagation algorithm is established for DRNN source separation. In the feedforward computation, the input observations $\mathbf{x}_t$ are composed of the magnitude spectrogram of the *mixing signal* at time frame $t$. The hidden layer features are calculated using Eq. (1) where $\mathbf{z}_t^{(0)} = \mathbf{x}_t$. The output layer consists of $\{\mathbf{y}_{1,t}(\mathbf{w}), \mathbf{y}_{2,t}(\mathbf{w})\}$ which are obtained via the predictive masking function by using the outputs of the last layer $L$ of DRNN $\{\mathbf{a}_{1,t}(\mathbf{w}), \mathbf{a}_{2,t}(\mathbf{w})\}$. Here, we choose the ideal radio mask [17] as the masking function

$$\mathbf{y}_{i,t}(\mathbf{w}) = \frac{|\mathbf{a}_{i,t}^{(L)}(\mathbf{w})|}{|\mathbf{a}_{1,t}^{(L)}(\mathbf{w})| + |\mathbf{a}_{2,t}^{(L)}(\mathbf{w})|} \quad (6)$$

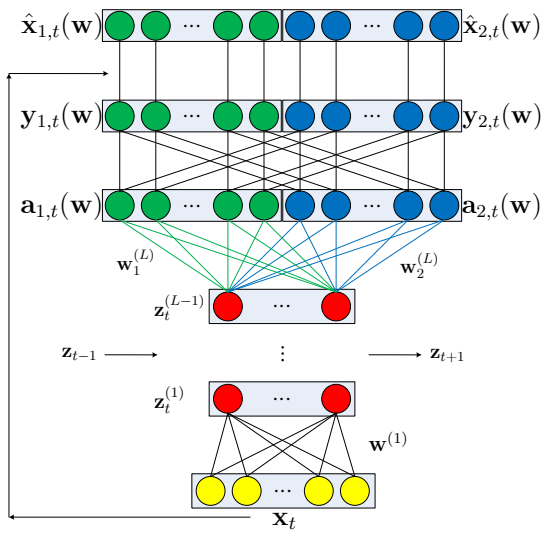


**Fig. 1**: Deep recurrent neural network for monaural source separation.

where $i \in \{1, 2\}$ represents the source index. The reconstructed magnitude spectrogram is calculated by multiplying the masking function over the mixed magnitude spectrogram $\mathbf{x}_t$ as:

$$\hat{\mathbf{x}}_{i,t}(\mathbf{w}) = \mathbf{x}_t \odot \mathbf{y}_{i,t}(\mathbf{w}) \tag{7}$$

where $\odot$ is the element-wise multiplication. The feedforward calculation of the other layers is the same as that in standard DRNN which is not shown here.

After the feedforward computation, the error backpropagation algorithm with stochastic gradient learning is developed so as to estimate the DRNN model parameters $\mathbf{w} = \{\mathbf{w}^{(l)}, \mathbf{w}^{(ll)}\}$. Instead of using the batch error function in Eq. (4), we adopt the error function which is calculated by using one data sample $\mathbf{x}_t$ in time $t$ as follows:

$$E_t(\mathbf{w}) = \frac{1}{2}\sum_{k=1}^{K}\Big\{(\hat{x}_{1,tk}(\mathbf{w}) - x_{1,tk})^2 + (\hat{x}_{2,tk}(\mathbf{w}) - x_{2,tk})^2 + \gamma(\hat{d}_{tk}(\mathbf{w}) - d_{tk})^2\Big\}. \tag{8}$$

To estimate the parameters corresponding to the *first* source signal, we derive the derivative of $E_t(\mathbf{w})$ with respect to the output-layer weights $w_{1,kj}^{(L)}$ which is obtained by using chain rule

$$\frac{\partial E_t(\mathbf{w})}{\partial w_{1,kj}^{(L)}} = \sum_{i=1}^{2}\left(\frac{\partial E_t(\mathbf{w})}{\partial \hat{x}_{i,tk}(\mathbf{w})}\frac{\partial \hat{x}_{i,tk}(\mathbf{w})}{\partial y_{i,tk}(\mathbf{w})}\frac{\partial y_{i,tk}(\mathbf{w})}{\partial a_{1,tk}^{(L)}(\mathbf{w})}\right)\frac{\partial a_{1,tk}^{(L)}(\mathbf{w})}{\partial w_{1,kj}^{(L)}} = \delta_{1,k}^{(L)} z_j^{(L-1)}. \tag{9}$$

Notably, the weight parameter $w_{1,kj}^{(L)}$ connects the hidden unit $j$ to the output unit $k$. The local gradient $\delta_{1,k}^{(L)}$ and the output of neuron at layer $L-1$, $z_j^{(L-1)}$, are defined for this derivative as follows

$$\delta_{1,k}^{(L)} = \sum_{i=1}^{2}\frac{\partial E_t(\mathbf{w})}{\partial \hat{x}_{i,tk}(\mathbf{w})}\frac{\partial \hat{x}_{i,tk}(\mathbf{w})}{\partial y_{i,tk}(\mathbf{w})}\frac{\partial y_{i,tk}(\mathbf{w})}{\partial a_{1,t,k}^{(L)}(\mathbf{w})} \tag{10}$$

$$z_j^{(L-1)} = \frac{\partial a_{1,tk}^{(L)}(\mathbf{w})}{\partial w_{1,kj}^{(L)}} \tag{11}$$

where the terms in Eq. (10) are yielded as

$$\frac{\partial E_t(\mathbf{w})}{\partial \hat{x}_{1,tk}(\mathbf{w})} = (1+\gamma)(\hat{x}_{1,tk}(\mathbf{w}) - x_{1,tk}) - \gamma(\hat{x}_{2,tk}(\mathbf{w}) - x_{2,tk}) \tag{12}$$

$$\frac{\partial E_t(\mathbf{w})}{\partial \hat{x}_{2,tk}(\mathbf{w})} = (1+\gamma)(\hat{x}_{2,tk}(\mathbf{w}) - x_{2,tk}) - \gamma(\hat{x}_{1,tk}(\mathbf{w}) - x_{1,tk}) \tag{13}$$

$$\frac{\partial \hat{x}_{1,tk}(\mathbf{w})}{\partial y_{1,tk}(\mathbf{w})} = \frac{\partial \hat{x}_{2,tk}(\mathbf{w})}{\partial y_{2,tk}(\mathbf{w})} = x_{tk} \tag{14}$$

$$\frac{\partial y_{1,tk}(\mathbf{w})}{\partial a_{1,tk}^{(L)}(\mathbf{w})} = \text{sgn}(a_{1,tk}^{(L)}(\mathbf{w}))\frac{y_{2,tk}(\mathbf{w})}{|a_{1,tk}^{(L)}(\mathbf{w})| + |a_{2,tk}^{(L)}(\mathbf{w})|} \tag{15}$$

$$\frac{\partial y_{2,tk}(\mathbf{w})}{\partial a_{1,tk}^{(L)}(\mathbf{w})} = -\text{sgn}(a_{1,tk}^{(L)}(\mathbf{w}))\frac{y_{2,tk}(\mathbf{w})}{|a_{1,tk}^{(L)}(\mathbf{w})| + |a_{2,tk}^{(L)}(\mathbf{w})|} \tag{16}$$

where the sgn extracts the sign of a real number. By combining Eqs. (12)-(16), we can derive the local gradient as

$$\delta_{1,k}^{(L)} = \Big\{(1+2\gamma)\Big[(\hat{x}_{1,tk}(\mathbf{w}) - x_{1,tk}) - (\hat{x}_{2,tk}(\mathbf{w}) - x_{2,tk})\Big]\Big\} \times x_{tk} y_{2,tk}(\mathbf{w})\frac{\text{sgn}(a_{1,tk}^{(L)}(\mathbf{w}))}{|a_{1,tk}^{(L)}(\mathbf{w})| + |a_{2,tk}^{(L)}(\mathbf{w})|}. \tag{17}$$

In a similar way, we can derive the derivative of $E_t(\mathbf{w})$ with respect to the output-layer weight for the *second* source, $\partial E_t(\mathbf{w})/\partial w_{2,kj}^{(L)}$, which is not shown in this paper.

Since we use DRNN model, the local gradient of $l$-th hidden layer units, $\delta_{t,j}^{(l)}$, for different layers $l \in \{1, \cdots, L-1\}$, can be also derived by applying the chain rule as

$$\begin{aligned}\delta_{t,j}^{(l)} &= \frac{\partial E_t(\mathbf{w})}{\partial a_{tj}^{(l)}} + \frac{\partial E_{t+1}(\mathbf{w})}{\partial a_{tj}^{(l)}} \\ &= \sum_k \frac{\partial E_t(\mathbf{w})}{\partial a_{tk}^{(l+1)}}\frac{\partial a_{tk}^{(l+1)}}{\partial a_{tk}^{(l)}} + \sum_k \frac{\partial E_{t+1}(\mathbf{w})}{\partial a_{t+1,k}^{(l)}}\frac{\partial a_{t+1,k}^{(l)}}{\partial a_{tk}^{(l)}} \\ &= f'(a_{tj}^{(l)})\left(\sum_k w_{kj}^{(l+1)}\delta_{tk}^{(l+1)} + \sum_k w_{kj}^{(ll)}\delta_{t+1,k}^{(l)}\right)\end{aligned} \tag{18}$$

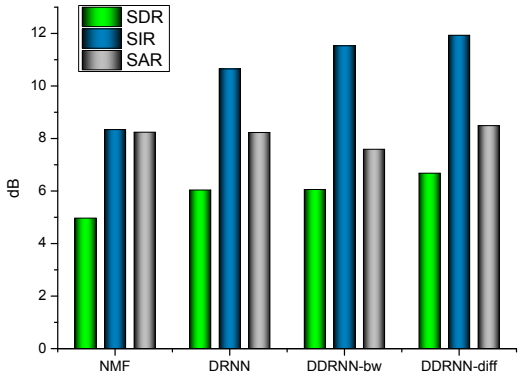


**Fig. 2**: Comparison of SDR, SIR and SAR of the separated signals using various separation methods

where $f(\cdot)$ is the activation function as used in Eq. (1) and $E_t(\mathbf{w})$ and $E_{t+1}(\mathbf{w})$ are the error functions calculated at times $t$ and $t+1$, respectively. The calculation of local gradient over two time steps is because of the recurrence in DRNN with forward weights $\mathbf{w}^{(l+1)}$ and recurrent weights $\mathbf{w}^{(ll)}$. Such errors are then propagated backwards through all layers to adjust the weights by using optimization methods. In this paper, we adopt the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method [18] in the optimization procedure for DRNN.

## 4. EXPERIMENTS

### 4.1. Experimental conditions

In the experiments, we use the mixed speech signals from TIMIT corpus [19] for evaluation of separation performance using different methods. There are 630 speakers in TIMIT corpus. Each speaker provides ten sentences. In training phase, eight sentences are randomly chosen from one male speaker and one female speaker for signal mixing. Another one sentence is used for cross validation and the remaining sentence is used for testing. All sentences are normalized to be with equal power. In order to rich the variety of the training samples, the sentences from one speaker are circularly shifted and added to the sentences from the other speaker as the training data. The 1024-point short-term Fourier transform with a 64-ms frame duration and a 32-ms frame shift is calculated to obtain the Fourier spectrograms. The spectra of the mixing speech are used as input features of the DRNN models. The DRNN architecture used in the experiments is fixed as 513-150-150-1026, which implies that the sizes are 513 for the input layer, 150 for two hidden layers, and 1026 (513*2) for the two source signals in output layer. The activation function ReLU is used in this study.

### 4.2. Experimental results

For comparative study, three objective functions are implemented in the same DRNN architecture. The objective function in Eq. (4) based on difference vectors is referred to as the discriminative DRNN (DDRNN)-diff and the objective functions in Eq. (2) and in Eq. (3) based on between-source information are referred to as DRNN and DDRNN-bw, respectively. The regularization parameter $\gamma$ in DDRNN-bw and DDRNN-diff are determined by validation data. The performance of NMF is carried out for comparison. The number of bases is determined by using validation data. The separation performance is assessed by using the source-to-distortion ratio (SDR), the source-to-interferences ratio (SIR), and the source-to-artifacts ratio (SAR) [20]. Figure 2 demonstrates the comparison of three metrics by using NMF and three DRNNs with different objective functions. There are some findings from this comparison. First, three DRNN algorithms perform better than the conventional NMF in terms of SDRs and SIRs. Comparing DRNN and DDRNN-bw, we find that DDRNN-bw outperforms DRNN in terms of SIR because that the additional between-source term is introduced in the objective function of DDRNN-bw in Eq. (3). This term is seen as a discrimination information which can reduce the interference between two source signals very well and accordingly improve the SIR of the demixed signals. Furthermore, the proposed method DDRNN-diff obtains higher SDRs and SIRs compared with NMF, DRNN and DDRNN-bw. SIR value of DDRNN-diff is comparable with that of DDRNN-bw. The discrimination measures using the difference vector between the constructed signals and the true signals helps the trained DRNN parameters for monaural source separation.

## 5. CONCLUSIONS

In this paper, we have presented a new objective function for discriminative learning for DRNN-based monaural source separation. The proposed objective function conveys the discrimination information which aims to preserve the mutual difference between two source signals during the supervised training procedure. This discrimination information is seen as a kind of regularization so as to maintain the between-source separation. An error backpropagation algorithm with a soft-masking function is developed to estimate the DRNN parameters in different forward layers and recurrent layers for separation of a mixed signal in presence of two source signals. The advantage of the proposed objective function is illustrated through the experiments on single-channel speech separation. It is shown that higher SDRs and SIRs are achieved by using the proposed discriminative DRNN when compared with NMF and DRNN using other objective functions. In the future, the discriminative objective function will be further explored by considering phase information and applied to train the general stochastic network (GSN) [21].

## 6. REFERENCES


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

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[3] J.-T. Chien and B.-C. Chen, "A new independent component analysis for speech recognition and separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1245–1254, 2006.

[4] J.-T. Chien and H.-L. Hsieh, "Convex divergence ICA for blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 302–313, 2012.

[5] J.-T. Chien and H.-L. Hsieh, "Nonstationary source separation using sequential and variational Bayesian learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 681–694, 2013.

[6] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. of Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2007, pp. 2614–2617.

[7] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[8] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564–1578, 2007.

[9] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 865–869.

[10] J.-T. Chien and P.-K. Yang, "Bayesian factorization and learning for monaural source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 1, pp. 185–195, 2016.

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

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

[13] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[14] H. Lee, P. Pham, Y. Largman, and A. Y Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1096–1104.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[16] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[17] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7092–7096.

[18] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. V. Le, and A. Y. Ng, "On optimization methods for deep learning," in *Proc. of International Conference on Machine Learning (ICML)*, 2011, pp. 265–272.

[19] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351 – 356, 1990.

[20] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[21] M. Zöhrer and F. Pernkopf, "General stochastic networks for classification," in *Advances in Neural Information Processing Systems*, 2014, pp. 2015–2023.