Monaural Source Separation Based on Adaptive Discriminative Criterion in Neural Networks

Yang Sun[1], Lei Zhu[2], Jonathon A. Chambers[1], Syed Mohsen Naqvi[1]
[1]School of Electrical and Electronic Engineering, Newcastle University, UK
[2]Science College, Harbin Engineering University, P.R. China
Email: y.sun29@newcastle.ac.uk

***Abstract*—Monaural source separation is an important research area which can help to improve the performance of several real-world applications, such as speech recognition and assisted living systems. Huang et al. proposed deep recurrent neural networks (DRNNs) with discriminative criterion objective function to improve the performance of source separation. However, the penalty factor in the objective function is selected randomly and empirically. Therefore, we introduce an approach to calculate the parameter in the discriminative term adaptively via the discrepancy between target features. The penalty factor can be changed with inputs to improve the separation performance. The proposed method is evaluated with different settings and architectures of neural networks. In these experiments, the TIMIT corpus is explored as the database and the signal to distortion ratio (SDR) as the measurement. Comparing with the previous approach, our method has improved robustness and a better separation performance.**

***Index Terms*—Monaural Source Separation, Deep Recurrent Neural Network, Penalty Factor, Adaptive**

I. Introduction

Speech source separation is a promising research topic for various real-world applications, such as automatic speech recognition (ASR), assisted living systems and hearing aids [1]–[3]. Some approaches have been utilized to single out sources from the speech mixtures by using spatial information and statistical properties of the speech signals, e.g. independent component analysis (ICA) and computational auditory scene analysis (CASA) [4]–[8]. While in the monaural source separation problem, only one speech mixture is captured and therefore the aforementioned methods become ill-posed. To solve the monaural source separation problem, several approaches have been proposed [9], [10]. One of the most famous methods is non-negative matrix factorization (NMF), which is a well established method for single channel speech separation [9]. However, because of the randomness in speech signals, the NMF based approaches are not expressive enough to model the complicated mapping function in many real-world scenarios [1].

In order to model highly non-linear mappings between the mixture and speech signals or a mixed signal to a time-frequency (T-F) mask, deep neural networks (DNNs) have been introduced [11]. In DNNs, the relationship can be obtained by optimizing the parameters of the networks. After the mapping function is learnt, the T-F masks or clean spectra are estimated and applied to reconstruct the desired speech signals. The T-F masks are categorized as binary masks or soft masks. In the binary mask, the T-F unit is assigned as 1 or 0 according to the criterion for the active source [12]. In the soft mask, the T-F unit is assigned as ratios of target energy and mixture energy [11]. In recent years, recurrent neural networks (RNNs) have provided state-of-art performance in speech signal processing, e.g. speech source separation, enhancement and recognition [1], [2], [13]. However, such RNNs often require high memory and computational power resource. In order to overcome these drawbacks of RNNs, the DRNNs are proposed, for which only the selected layers in the networks have the temporal connection [1].

In this paper, the DRNNs are trained to estimate the binary and soft T-F masks. Different architectures of DRNNs are used to generate the T-F masks to separate the speech mixture with the discriminative training criterion. In our method, the parameter in the discriminative term is calculated adaptively to penalize the objective function. The DRNNs with the proposed adaptive discriminative criterion outperform the performance of [1].

The remainder of the paper is organized as follows, in Section II, the architectures of DRNNs and the T-F masks are described. In Section III, the method to calculate the penalty factor in the discriminative term with different norms is presented; experimental settings and results are shown in Section IV to confirm the improvement of the proposed approach. Finally, conclusions are drawn in Section V.

II. Relation to previous work

1) Architectures of Neural Networks: In the monaural source separation problem, which is solved via neural networks, the separation performance can be improved by utilizing the temporal information of the speech signals in the training stage of networks. Commonly, the temporal information is exploited in two ways: concatenating neighbouring features and using RNNs [14]. In the concatenating features method, a larger window size can utilize more temporal information with the trade off being computational and memory resources. Therefore, an appropriate window size is required. The RNNs have a recurrent architecture, which is a powerful model for temporal information. The DRNNs combine the multiple

This work was supported by the China Scholarship Council (CSC) Grant number 201606685042.

levels of representation that have proved so effective in DNNs with the flexible use of long range context that empowers RNNs [13].

According to [15], two DRNN architectures are defined: 1) an L hidden layer DRNN with temporal connection only at the l-th layer (DRNN-l) and 2) a full RNN. Assume $\mathbf{h}_t^l$ is the hidden activation at layer l and time t :

$$\begin{aligned}\mathbf{h}_t^l &= f_{\mathrm{h}}(\mathbf{x}_t, \mathbf{h}_{t-1}^l) \\ &= g_l(\mathbf{R}^l\mathbf{h}_{t-1}^l + \mathbf{W}^l g_{l-1}(\mathbf{W}^{l-1}(\cdots g_1(\mathbf{W}^1\mathbf{x}_{\boldsymbol{t}})))) \end{aligned} \tag{1}$$

the output $\mathbf{y}_t$ is expressed as:

$$\begin{aligned}\mathbf{y}_t &= f_o(\mathbf{h}_t^l) \\ &= \mathbf{W}^L g_{L-1}(\mathbf{W}^{L-1}(\cdots g_l(\mathbf{W}^l\mathbf{h}_t^l)))\end{aligned} \tag{2}$$

where f_h and f_o are the state transition and output function, respectively. The input at time t is $\mathbf{x}_t$, $g(\cdot)_l$ represents the activation function at the l-th layer, $\mathbf{R}^l$ is the recurrent weight matrix and $\mathbf{W}^l$ is the current connection at the l-th layer. In the layers without temporal connection, the previous weight matrices are the zero matrices.

The full connection DRNN has the same architecture as the vanilla RNN [16], the hidden state of the l-th layer at time t is:

$$\mathbf{h}_t^l = f_{\mathrm{h}}(\mathbf{h}_t^{l-1}, \mathbf{h}_{t-1}^l) = g_l(\mathbf{R}^l\mathbf{h}_{t-1}^l + \mathbf{W}^l\mathbf{h}_t^{l-1}) \tag{3}$$

In the first layer, where l = 1, the activation $\mathbf{h}_t^1$ is calculated by $\mathbf{h}_t^0 = \mathbf{x}_t$. In the DRNN, the activation function is selected as a rectified linear unit (ReLU) to avoid gradient vanishing and reduce the computational cost. The ReLU function is expressed as:

$$g(\mathbf{x}) = max(\mathbf{0}, \mathbf{x}) \tag{4}$$

2) Time-Frequency Mask: The proposed method trains the neural network to learn the mapping relationship from the features of the mixed signal to the features of the source signals and the T-F mask is computed by using the output features. In this work, both the binary and soft masking functions are explored.

Assume the target outputs of the neural networks are $\mathbf{y}_{1t}$ and $\mathbf{y}_{2t}$, the predicted outputs are $\hat{\mathbf{y}}_{1t}$ and $\hat{\mathbf{y}}_{2t}$. These outputs are the magnitude spectra of source 1 and 2.

Using the estimated outputs to generate masks to separate signals, the binary T-F mask is computed as [17]:

$$\mathbf{M}_{\boldsymbol{b}}(f,t) = \begin{cases} 1 & |\hat{\mathbf{y}}_{1t}(f)| > |\hat{\mathbf{y}}_{2t}(f)| \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The soft T-F mask for source 1 is expressed as [18]:

$$\mathbf{M}_{\boldsymbol{s}}(f,t) = \frac{|\hat{\mathbf{y}}_{1t}(f)|}{|\hat{\mathbf{y}}_{1t}(f)| + |\hat{\mathbf{y}}_{2t}(f)|} \tag{6}$$

where $f = 1, \cdots F$ is the index of the frequency bins, $t = 1, \cdots T$ is the index of the temporal frame bins.

Assume $\mathbf{X}_t$ is the magnitude spectra of the input mixture signal, the separated spectra can be computed as:

$$\hat{\mathbf{s}}_{1t} = \mathbf{M}\odot\mathbf{X}_t$$

$$\hat{\mathbf{s}}_{2t} = (1 - \mathbf{M})\odot\mathbf{X}_t \tag{7}$$

where the mask $\mathbf{M}$ can be selected as a binary or soft mask, $\odot$ is the element-wise multiplication operator. By using the inverse short-time Fourier transform (ISTFT), the source speech signals are reconstructed.

Because well-trained neural networks provide more accurate estimated spectra, they can help to improve the separation performance. Based on these network architectures and T-F masks, our proposed approach will focus on how to optimize the neural network parameters. The penalty parameter in the objective function is calculated adaptively, which will be elaborated in the next section.

III. Proposed Method

By optimizing the parameters of the neural network, the mapping relationship between the feature of mixture, $\mathbf{x}_t$, and the estimations, $\hat{\mathbf{y}}_{1t}$ and $\hat{\mathbf{y}}_{2t}$, can be obtained. The sum of the squared errors is selected as the objective function as:

$$J = \frac{1}{2}\sum_{t=1}^{T}(\|\hat{\mathbf{y}}_{1t} - \mathbf{y}_{1t}\|_2^2 + \|\hat{\mathbf{y}}_{2t} - \mathbf{y}_{2t}\|_2^2) \tag{8}$$

where $\hat{\mathbf{y}}_{1t}$ and $\hat{\mathbf{y}}_{2t}$ are the predictions of the spectra and $\mathbf{y}_{1t}$ and $\mathbf{y}_{2t}$ represent the target spectra, $\|\cdot\|_2^2$ is the l_2 norm operation, and (8) needs to be minimized to optimize the parameters in the neural network.

In this work, the input is a concatenation of features; when the features are similar, the neural network will be conservative in the training stage. Because of the similarity, a feature can be attributed to source 1 or source 2 in some cases. To maintain the efficiency of the training stage, the neural network will attribute the feature to both source 1 and source 2, which is called the conservative strategy. However, if the ambiguous features are attributed repeatedly, the separation performance is decreased due to this strategy.

In [1], a discriminative network training criterion was proposed. The new discriminative objective function is defined as:

$$\begin{aligned}J_{DIS} = \frac{1}{2}\sum_{t=1}^{T}(&\|\mathbf{y}_{1t} - \hat{\mathbf{y}}_{1t}\|^2 + \|\mathbf{y}_{2t} - \hat{\mathbf{y}}_{2t}\|^2 - \\ &\gamma\|\mathbf{y}_{1t} - \hat{\mathbf{y}}_{2t}\|^2 - \gamma\|\mathbf{y}_{2t} - \hat{\mathbf{y}}_{1t}\|^2)\end{aligned} \tag{9}$$

where γ can be treated as the penalty parameter. In the ideal case, $\hat{\mathbf{y}}_{1t}$ and $\hat{\mathbf{y}}_{2t}$ are only estimated by the corresponding target features. However, because of the indeterminacy and conservative strategy, this case cannot happen. What we can do is to minimize the negative influence from these ambiguous features. The $\|\mathbf{y}_{1t} - \hat{\mathbf{y}}_{2t}\|^2$ and $\|\mathbf{y}_{2t} - \hat{\mathbf{y}}_{1t}\|^2$ terms are used to represent the squared errors, which are caused by attributing the estimated features, $\hat{\mathbf{y}}_{1t}$ and $\hat{\mathbf{y}}_{2t}$, incorrectly.

According to previous work [1], γ is selected in the range of 0.01∼0.1, empirically. Whereas the speech signals are random with high indeterminacy. If the value of γ is irrelevant to inputs, when the inputs for training stage are changed, the performance and the trained network may not be amenable.

Therefore, we propose an approach to calculate the penalty parameter adaptively, which is applied to penalize the objective function to train the neural networks.

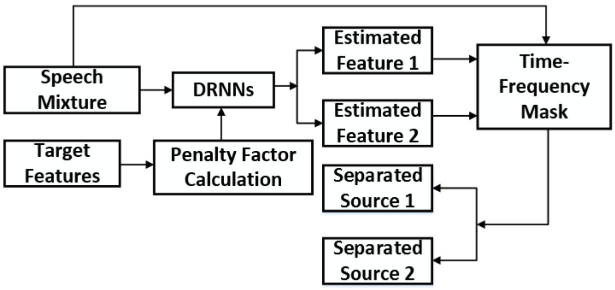

Fig. 1: Framework of the Proposed Method

Figure 1 is the flow diagram of our proposed method. Before training the neural network, a penalty factor calculation module is added to compute the parameter in the discriminative term to penalize the objective function. Then, in the training stage, the parameters of the DRNN are optimized with the penalty factor and discriminative criterion.

In our method, the value of γ in (9) is changed with the input features. To be specific, if the input features are almost the same, it indicates that features are more likely to be attributed to both source 1 and source 2. Therefore, the penalty term needs to be significant and the γ requires a greater value. In contrast, when the targets have huge differences, the conservative strategy and penalty factor are trivial in this situation and γ should be close to zero. According to the analysis above, the value of the penalty factor is inversely proportional to the discrepancy between target features.

Generally, norms of matrix are used to measure the discrepancy. In this paper, we explore three types of norms.

Assume the spectra of source 1 and source 2 are, respectively, $\mathbf{A} \in \mathbb{R}^{F\times T}$ and $\mathbf{B} \in \mathbb{R}^{F\times T}$. The discrepancy between the features is defined as:

$$\mathbf{D} = \mathbf{A} - \mathbf{B} \tag{10}$$

The penalty factor is calculated as:

$$\gamma = \frac{1}{\|\mathbf{D}\|_{norm}} \tag{11}$$

Because the discrepancy between two features needs to be measured, firstly, the max norm is utilized, which is defined as:

$$\|\mathbf{D}\|_{max} = \max|d_{t,f}| \qquad \forall\ t, f \tag{12}$$

where $d_{t,f}$ is the element in the matrix $\mathbf{D}$, t and f represent the frame and frequency index: $t = 1, \ldots, T$ and $f = 1, \ldots, F$.

However, the max norm only finds the maximum value of the matrix, it cannot fully measure the total discrepancy. Hence, the P-norm will be discussed below.

The P-norm of matrix $\mathbf{D}$ is defined as:

$$\|\mathbf{D}\|_P = (\sum_{t=1}^{T}\sum_{f=1}^{F}|d_{t,f}|^P)^{\frac{1}{P}} \tag{13}$$

where P is the positive integer.

In this work, we discuss two cases in the P-norm, where the value of P is selected as 1 or 2.
For P = 1:

$$\|\mathbf{D}\|_1 = \sum_{t=1}^{T}\sum_{f=1}^{F}|d_{t,f}| \tag{14}$$

For P = 2:

$$\|\mathbf{D}\|_2 = (\sum_{t=1}^{T}\sum_{f=1}^{F}|d_{t,f}|^2)^{\frac{1}{2}} = \sqrt{trace(\mathbf{D}{\cdot}\mathbf{D}^*)} \tag{15}$$

where $\mathbf{D}^*$ denotes the conjugate transpose of $\mathbf{D}$. It is well known as the Frobenius norm.

Theoretically, from the definition of the 2-norm, we can know that it shrinks the difference between inputs. Therefore, the algorithm based on the 1-norm should have a better separation performance. To confirm this point, the performance of DNNs with different P-norms is compared in Table 1.

TABLE 1: Separation performance comparison in terms of SDR (dB) with different types of P-norm, the architecture of these neural networks are DNNs.

Norm Types	1-norm	2-norm
Binary Mask	6.64	6.28
Soft Mask	7.27	6.92

Moreover, for any two matrix norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, they have the relationship for some positive constants δ and θ and all matrices $\mathbf{D}$ in $\mathbb{R}^{F\times T}$. It is defined as:

$$\delta\|\mathbf{D}\|_\alpha \leqslant \|\mathbf{D}\|_\beta \leqslant \theta\|\mathbf{D}\|_\alpha \tag{16}$$

The above equation indicates that all norms on $\mathbb{R}^{F\times T}$ are equivalent [19]. However, in a specific algorithm, the 1-norm and the 2-norm will show different performance. From Table 1, the 1-norm is the proper choice.

Finally, the type of norm in (11) is selected as the 1-norm and the penalty factor is calculated as:

$$\gamma = \frac{1}{\|\mathbf{D}\|_1} = \frac{1}{\|\mathbf{A} - \mathbf{B}\|_1} \tag{17}$$

Therefore, the γ can be calculated adaptively with the changes of target features.

This approach is effective for all of the neural network architectures in Section II and considers both interpretability and precision of the discriminative parameter, which will be confirmed by experimental results in the next section.

IV. Experiments

1) Experimental Settings: The separation performance is evaluated based on the famous TIMIT database, which contains broadband recordings of 630 speakers [20]. In our experiments, speech signals are selected from the TIMIT corpus randomly to constitute the training, validation and testing sets. The number of mixtures in training, validation and testing set is 972, 216 and 108, respectively. The mixtures in these experiments are generated with different speech sources having different genders. To extract the proper spectral representation to train the networks, a 1024-point short time

Fourier transform (STFT) with 50 % overlap is explored. The initialization method in [21] is utilized to reduce the training difficulty of deep networks.

The circular shift in the time domain is explored to increase the variety of training set [22]. The spectra and log power spectra are utilized as the types of input features, which are calculated by using the HTK toolkit [23]. The basic DNN, the DRNN with first layer connection, the DRNN with second layer connection and full connected DRNN are the four different architectures of neural networks. All of experiments are based on these architectures to identify generalization ability of the proposed method.

In these networks, the number of hidden layers is two and the number of hidden units on each layer is 1000. The SDR is utilized to measure the separation performance of the proposed method [24]. The limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method is an optimization algorithm in the family of quasi-Newton methods, which is used to train the models [25]. In the experiments, the values of γ are selected as 0, 1 and 0.05 (in the range of 0.01 and 0.1) for comparison. The size of context window in these networks is 1, the concatenation contains three frames, one central frame and two window frames. According to the analysis in Section III and Table 1, the 1-norm is applied to calculate the discrepancy in the target features.

2) Experimental Results: After the different neural networks are trained, the mixture is separated by using different mask functions and values of γ.

TABLE 2: Separation performance comparison in terms of SDR (dB) with different values of γ and neural network architectures via binary mask and the input features are spectra.

Penalty factor γ	DNN	DRNN-1	DRNN-2	RNN
γ = 1	5.49	5.61	6.60	6.56
γ = 0	5.25	5.38	6.57	5.91
γ = 0.05 [1]	5.50	5.58	6.52	6.72
Adaptive γ	5.81	5.96	6.66	6.84

TABLE 3: Separation performance comparison in terms of SDR (dB) with different values of γ and neural network architectures via binary mask and the input features are log power spectra.

Penalty factor γ	DNN	DRNN-1	DRNN-2	RNN
γ = 1	5.56	5.89	6.12	6.62
γ = 0	5.13	6.16	5.88	7.01
γ = 0.05 [1]	6.28	6.56	6.27	6.94
Adaptive γ	6.79	6.89	6.87	7.11

TABLE 4: Separation performance comparison in terms of SDR (dB) with different values of γ and neural network architectures via soft mask and the input features are spectra.

Penalty factor γ	DNN	DRNN-1	DRNN-2	RNN
γ = 1	6.08	6.17	7.00	7.07
γ = 0	5.72	6.20	7.12	6.60
γ = 0.05 [1]	6.14	6.25	7.24	7.52
Adaptive γ	6.30	6.70	7.48	7.56

TABLE 5: Separation performance comparison in terms of SDR (dB) with different values of γ and neural network architectures via soft mask and the input features are log power spectra.

Penalty factor γ	DNN	DRNN-1	DRNN-2	RNN
γ = 1	6.01	6.13	6.75	7.21
γ = 0	6.13	6.51	6.31	6.77
γ = 0.05 [1]	6.82	7.23	7.26	7.40
Adaptive γ	7.07	7.52	7.33	7.74

The experimental results are compared in terms of different aspects. Firstly, it can be seen from Tables 2 & 3 and Tables 4 & 5 that the separation performance is impacted by the types of features in different architectures of networks. Generally, in DNN and DRNN-1, using the log power spectra as the input features has better performance. In contrast, the spectra can yield a higher SDR in DRNN-2 and full RNN. Then, according to the Tables 2 & 4 and Tables 3 & 5, the soft mask based models outperform binary mask based models greatly. It is evident that the soft mask can have around 10% more improvements in SDR.

Finally, the performance between different architectures is compared. The results in all Tables confirm the separation performance and robustness of the proposed method are improved in all architectures of DRNNs. Besides, comparing the separation performance of DNN and DRNNs, introducing the connected layer in networks can provide improvement. In DRNNs, almost all of the full RNN maintains the highest SDR, but demands high computational power and larger memory. In these architectures with connection in hidden layers, DRNN-1, DRNN-2 and full RNN, increasing the complexities of DRNNs gains the SDR. Although the performance is affected differently for DNN and DRNNs, the proposed approach outperforms the DRNN-based method in [1].

In the experiments, the proposed method is compared with different architectures and values of penalty factors. From Table 1, the 1-norm is the proper choice to calculate the penalty factor. According to Tables 2-5, the results of the proposed method surpass the experimental results, which are produced by the irrelevant parameter method. The soft masking function can assist to achieve a better separation performance. Generally, the full RNN is the better choice than DNN, DRNN-1 and DRNN-2, but the requirement of computational resource will be higher, when the complexity of the network is increased.

V. Conclusion

In this paper, we proposed a method for learning with an adaptive penalty factor. DRNNs were trained by our proposed approach to solve the monaural speech source separation problem. Various neural network architectures and different values of γ were explored based on our approach. All of the experimental results confirmed that the adaptive criterion method outperformed the approach with irrelevant penalty factor method [1]. Because of the indeterminacy of speech signals in the real-world scenarios, our method can be more applicable. In the future work, we will explore some new inverse proportional functions or limited numerical precision DRNNs to further improve the separation performance and efficiency.

References

[1] P. Huang, M. Kim, M.-H. 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, 2015.

[2] B. Wu, K. Li, M. Yang, and C.-H. Lee, “ A Reverberation Time Aware Approach to Speech Dereverberation Based on Deep Neural Networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, 2017.

[3] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang, and J. A. Chambers, “A Posture recognition-based fall detection system for monitoring an elderly person in a smart home environment,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1274–1286, 2012.

[4] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, “Audiovisual speech source separation: An overview of key methodologies,” *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014.

[5] S. M. Naqvi, M. Yu, and J. A. Chambers, “A Multimodal Approach to Blind Source Separation of Moving Sources,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 895–910, 2010.

[6] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, “Model-based expectation-maximization source separation and localization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[7] M. S. Salman, S. M. Naqvi, A. Rehman, W. Wang, and J. A. Chambers, “Video-Aided Model-Based Source Separation in Real Reverberant Rooms,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1900–1912, 2013.

[8] Y. Sun, W. Rafique, J. A. Chambers, and S. M. Naqvi, “Underdetermined source separation using time-frequency masks and an adaptive combined gaussian-students t probabilistic model,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[9] D. D. Lee and H. S. Seung, “ Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[10] P. Smaragdis, B. Raj, and M. Shashanka, “A probabilistic latent variable model for acoustic modelling,” in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2006.

[11] X. L. Zhang and D. L. Wang, “ A Deep Ensemble Learining Method for Monaural Speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 967–977, 2016.

[12] D. L. Wang, “ On Ideal Binary mask as the computational goal of auditory scene analysis,” *Speech Sep. Humans Mach*, vol. 60, pp. 63–64, 2005.

[13] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[14] Y. Wang, J. Du, L. R. Dai, and C.-H. Lee, “Unsupervised single-channel speech separation via deep neural network for different gender,” in *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016.

[15] M. Hermans and B. Schrauwen, “Training and analysing deep recurrent neural networks,” in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2013.

[16] A. C. I. Goodfellow and Y. Bengio, *Deep Learning*. MIT Press, 2016.

[17] N. R. S. Srinivasan and D. L. Wang, “Binary and ratio time-frequency masks for robust speech recognition,” *Speech Communication*, vol. 48, no. 11, p. 14861501, 2006.

[18] E. Ceolini and S.-C. Liu, “Impact of low-precision deep regression networks on single-channel source separation,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[19] J. W. Demmel, *Applied Numerical Linear Algebra*. Siam, 1997.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, 1993.

[21] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proc. Conf. Artificial Intelligence and Statistics*, 2010.

[22] P. Huang, M. Kim, M.-H. Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[23] *Hidden Markov Model Toolkit (HTK)*. Cambridge University and Microsoft, 2016.

[24] 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, 2006.

[25] R. Fletcher, *Practical Methods of Optimization, 2nd Edition*. Wiley, 2000.