ISSN: 1803-3814 (Printed), 2571-3701 (Online)
https://doi.org/10.13164/mendel.2021.2.044

Hybrid Deep Learning Model for Singing Voice Separation

Rusul Amer✉, **Ahmed Al Tmeme**✉

Information and Communication Eng. Dept., Al Khwarizmi Eng. College, University of Baghdad, Iraq

rusul.a.barrak@gmail.com ✉, Asattar@kecbu.uobaghdad.edu.iq✉

Abstract

Monaural source separation is a challenging issue due to the fact that there is only a single channel available; however, there is an unlimited range of possible solutions. In this paper, a monaural source separation model based hybrid deep learning model, which consists of convolution neural network (CNN), dense neural network (DNN) and recurrent neural network (RNN), will be presented. A trial and error method will be used to optimize the number of layers in the proposed model. Moreover, the effects of the learning rate, optimization algorithms, and the number of epochs on the separation performance will be explored. Our model was evaluated using the MIR-1K dataset for singing voice separation. Moreover, the proposed approach achieves (4.81) dB GNSDR gain, (7.28) dB GSIR gain, and (3.39) dB GSAR gain in comparison to current approaches.

Keywords: *Monaural Source Separation, Hybrid Deep Learning, Time Frequency Masking, Convolution Neural Network, Dense Neural Network, Recurrent Neural Network.*

Received: 26 September 2021
Accepted: 10 December 2021
Published: 21 December 2021

1 Introduction

Source separation is the task of recovering the individual sources from a mixture signal [9]. Although the human auditory system has a remarkable capability in separating sounds originating from different sources, which is considered an effortless task for humans, it is difficult for machines. Monaural source separation, in particular, is considered more difficult as one channel is provided [7]. Therefore, it is still an open problem that brings the attention of researchers [20]. Speech separation, singing voice separation, and speech denoising are all examples of real-world applications where source separation is important.

Researchers have recently become interested in source separation-based deep learning algorithms, which model the nonlinear mapping connection between mixed and separated data [8], where it was suggested to use a deep neural network to compute Ideal Binary Masks (IBMs) that used for separating the speech signals from a noisy mixture [26]. Due to the fact that singing voice is one of the typical time-series signals, the RNN's internal state is utilized for capturing the behavior of the dynamic timing of those signals [15]. Therefore, a deep recurrent neural network (DRNN) [5] that has been generated through the stacking of the RNN is capable of effectively exploring the distribution of the information on various time scales for the music source voice separation (MSVS) [10]. In addition, convolution neural networks (CNN) have recently been utilized to extract vocals from a musical mixture [18]. Due to the fact that the CNN makes use of the tiny scale features present in data [12], as well as it has the ability to extract the translationally invariant and highly discriminative features [1]. Thus, it may be incorporated for assisting the RNN for the extraction of de-redundant and low-dimensional magnitude spectra representations.

The deep learning model is trained with optimization algorithms that are used to change the attributes of the model, such as weights and learning rate, in order to reduce the error between the predicted sources and the original sources. Adam optimizer [11] is the best algorithm as it combines the advantages of two methods: AdaGrad [4], which works well with sparse gradients, and Adadelta [31], which performs well in online and non-stationary scenarios, where adaptive learning rates for each parameter are calculated. Additionally to storing an exponentially decaying average of previous squared gradients, it also stores an exponentially decaying average of the previous gradient.

In this paper, a hybrid deep learning model that combines (CNN, DNN, and RNN) for the separation of the singing voice from the monaural recordings in a supervised manner that is jointly optimized with soft time-frequency masking has been proposed. We also propose a trial and error method to optimize the structure of our hybrid deep learning model. Furthermore, various training objectives will be investigated in order to optimize the model.

This research is created as follows: Section two provides a brief description of the related work. The proposed hybrid deep learning model and soft time-frequency masking function are presented in Section three. Section four discusses the experiments and results obtained using the MIR-1K dataset. Furthermore, the conclusions will be drawn in Section five.

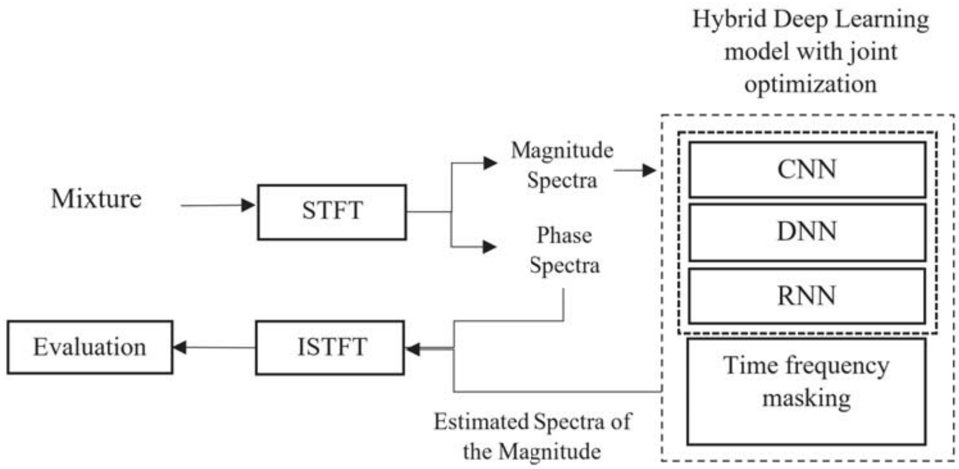


Figure 1: The proposed framework.

2 Relation to Previous Work

A wide variety of methods were proposed for source separation, Huang et al. [7] introduced a Robust Principal Component Analysis (RPCA) approach for solving underlying low-rank matrices, including music accompaniment and sparse matrices containing singing voice. Y.-H. Yang [28] proposed two modifications to the decomposition by (RPCA) that take harmonic similarity between sinusoids into consideration and that the use of drum removal after RPCA results in significant improvement. Y.-H. Yang [29] employed dictionary learning methods to estimate the subspace structures of musical sources and introduced a new approach called multiple low-rank representations (MLRR) that is used for decomposition by the learnt dictionaries. Al Tmem et al. [2, 3, 22, 27] proposed an algorithm based on the factorization method with a hybrid framework that combined both the multiplicative update and the Expectation-Maximization algorithms. Wang et al. [26] learned the ideal binary mask using deep neural networks, and source separation issues were regarded as binary classification problems. Likewise, Uhlich et al. [23] performed the extraction of instrument signals from the music with the use of the DNNs. Nugaraha et al. [14] utilized the DNNs for learning signal source spectral features and utilized the Wiener filters for distinguishing the signals from the noise. Huang et al. [9] employed a recurrent neural network on speech separation to learn from past time steps and get extended context information. Uhlich et al. [24] separated music sources using data augmentation and network integration. Sebastian et al. [16] learned the source's time-frequency mask using the modified group delay (MOD-GD) function. Sun et al. [21] have suggested a 2-stage method with 2 DNN-based approaches for addressing the issue of the efficiency of the current approaches of speech separation. CNN was commonly utilized in the area of deep learning, and presently it was implemented for the tasks of speech separation [17, 19], where it outperformed DNN-based speech separation systems and achieved optimal separation performance under the same condition. For tackling time-frequency masking challenges, Luo et al. [13] presented the Conv-TasNet, an entirely convolutional time-domain source separation network. Yuan et al. [30] have suggested an Enhanced Feature Network (EFN) that was capable of achieving a specific level of enhancement in GSAR as well as GNSDR indicators in comparison with DRNN.

The proposed model's advantage is represented by combining the CNN and the RNN for extracting effective de-redundant and low-dimensional representations from mixture signal magnitude spectra to avoid spectra decomposition high cost's, reducing training time in comparison to the DRNN, and separation performance that outperforms other approaches.

3 Proposed Methods

In this paper, we propose a hybrid deep learning model for source separation that utilizes three kinds of neural networks (CNN, DNN, and RNN). Fig. 1 illustrates the proposed framework. The mixture's spectrogram can be computed using the short time Fourier transform (STFT); the spectral of the magnitude is passed through the hybrid deep learning model in order to separate the mixture signal and produce an estimation for every separated source. The estimated sources are utilized to compute time-frequency soft masks, which are then used to isolate sources final magnitude estimations. In addition, those estimations, as well as the mixture phase, have been utilized by the inverse short-time Fourier transform (ISTFT) for obtaining audio signals that correspond to the sources that have been separated.

The propose framework can be explained as follows:

3.1 Hybrid Deep Learning Model Architecture

We propose to model the temporal information of audio signal by using a hybrid deep learning model that

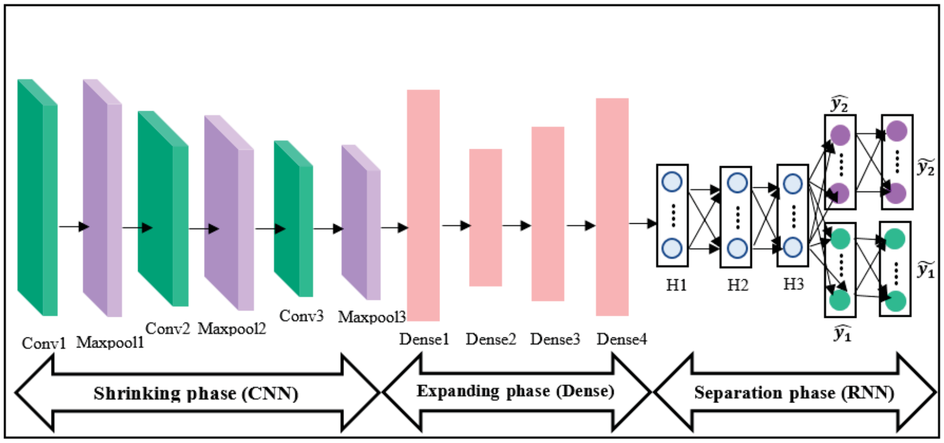


Figure 2: The proposed architecture of hybrid deep learning model.

Table 1: Parameter description of the hybrid deep learning model.

Layer	Filter/unit size	Kernel Size	Stride	Activation	Padding	Parameters
Conv 1D	512	3	1	ReLU	Same	2048
Max-pooling 1D	-	3	1	ReLU	-	-
Conv 1D	256	3	1	ReLU	Same	393472
Max-pooling 1D	-	3	1	ReLU	-	-
Conv 1D	128	3	1	ReLU	Same	98432
Max-pooling 1D	-	3	1	ReLU	-	-
Dense 1	1024	-	-	ReLU	-	132096
Dense 2	128	-	-	ReLU	-	131200
Dense 3	256	-	-	ReLU	-	33024
Dense 4	512	-	-	ReLU	-	131584
3 RNN	256	-	-	ReLU	-	3935232
Total trainable parameters						**4857088**

consists of three phases to learn how to reconstruct the target spectra, as follows:

The shrinking phase; represented by Convolution neural network (CNN): CNNs is a recently utilized approach for speech separation. Depending on the input data used in this paper, which is a one-dimensional time series audio signal, we use one-dimension CNN which is a modified version of the two-dimension CNN for extracting the effective de-redundant and low-dimensional representations from mixture signal magnitude spectra. This phase contain three convolutional layers and three max-pooling layers, the convolution layers with filter size (512, 256, 128, respectively) kernel 3, stride 1, padding 'SAME' accordingly, the feature map size doesn't change after convolution and rectified linear unit (ReLU) activation function. $f(x) = \max(0, x)$ which performs more sufficiently in comparison with the use of a sigmoid or tanh function. Max pooling layers after each convolution layer are used to reduce the dimensionality of the feature map.

The expanding phase; represented by Dense layers: this phase consists of four dense layers with units (1024, 128, 256, 512, respectively) used for collecting the extracted features from the previous phase.

The separation phase (reconstruction phase); represented by a RNN: this phase consists of three RNN layers with 256 hidden units, which is used for the extraction of high-level sequential feature from extracted features by the previous phases. The RNN function is fundamentally for using feature information that has been learned by convolutional layers for the separation of the vocals and accompaniments. Therefore, a CNN has been presented as RNN front-end for the purpose of extracting the global features and speech spectrogram fine details, like the harmonics. The backend used RNN, which includes a method for "memorizing" the sequenced data.

The proposed hybrid deep learning model is shown in Fig. 2.

3.2 Time-Frequency Masking

The time-frequency masking function ensures the total of the predicted results is identical to the mixture in its original state. Additionally, we noticed that using a time-frequency masking method to smooth the source separation findings is beneficial. We jointly train the model rather than individual training and then use the time-frequency masking to the output. Following [9, 10], we incorporate the computation of the soft masks as an additional deterministic layer into the network architecture as follows:

$$\widetilde{y}_{1t} = \frac{|\hat{y}_{1t}|}{|\hat{y}_{1t}| + |\hat{y}_{2t}|} \odot X_t \tag{1}$$

$$\widetilde{y}_{2t} = \frac{|\hat{y}_{2t}|}{|\hat{y}_{1t}| + |\hat{y}_{2t}|} \odot X_t \tag{2}$$

Where $\hat{y}_{1t}$ and $\hat{y}_{2t}$ the predictions through the network, X_t is the mixture's spectrum and the operator $\odot$ represents element-wise multiplication (Hadamard product). Thus, the network can be optimized with the masking function jointly [25].

3.3 Training Objectives

Considering output predictions $\hat{y}_{1t}$ & $\hat{y}_{2t}$ (or $\widetilde{y}_{1t}$ & $\widetilde{y}_{2t}$) of original sources y_{1t} and y_{2t}, the optimization of this model parameters has been explored through the minimization of squared error criteria, based on the following equation:

$$J_{MSE} = ||\hat{y}_{1t} - y_{1t}||_2^2 + ||\hat{y}_{2t} - y_{2t}||_2^2 \tag{3}$$

4 Experimental Results

4.1 Dataset

The MIR-1K dataset is used to evaluate our system [6]. One thousand music clips ranging in length from 4 to 13 seconds are encoded at a 16 kHz sampling rate. The clips were taken from 110 Chinese karaoke songs sung by 19 amateur singers (8 females and 11 males). We use 700 clips as the training, and 300 are used for testing.

4.2 Evaluation

The performance of the source separation can be measured by three quantitative values of the BSS-EVAL metrics: Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR), and Source to Artifacts Ratio (SAR). In addition, the Normalized SDR (NSDR) has been represented by:

$$NSDR(\hat{y}, y, x) = SDR(\hat{y}, y) - SDR(x, y) \tag{4}$$

where $\hat{y}$ represents the re-synthesized singing voice (vocals), y represents the clean vocals, and x represents the mixture. The NSDR is used to calculate the difference in SDR between a pre-processed mixture x and a separated vocals $\hat{y}$. Global SIR (GSIR), Global SAR (GSAR), and Global NSDR (GNSDR) values have been reported to indicate weighted mean values of SIRs, SARs, and NSDRs, respectively, of the test clips that have been weighted by length. Higher SIR, SDR, and SAR values denote a sufficient quality of the separation.

4.3 Experiments

In the experiments, the magnitude spectra feature has been used as input to the model. The spectral representation is obtained using a 1024-point Short Time Fourier transform (STFT) with a 25% overlap. The experiments can be divided into 4 phases:

Phase one: The trial and error method was proposed to estimate the number of layers in each phase until the best source separation performance was reached. Fig. 3 shows that, by using 3CNN layers, 4 DNN layers and 3 RNN layers, the model provide higher GNSDRs, GSIRs and GSARs.

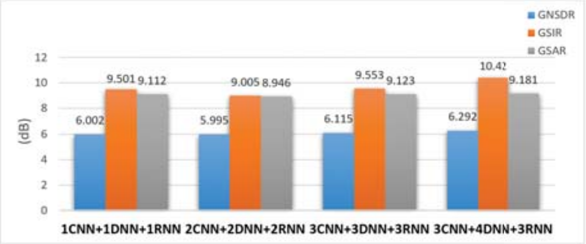
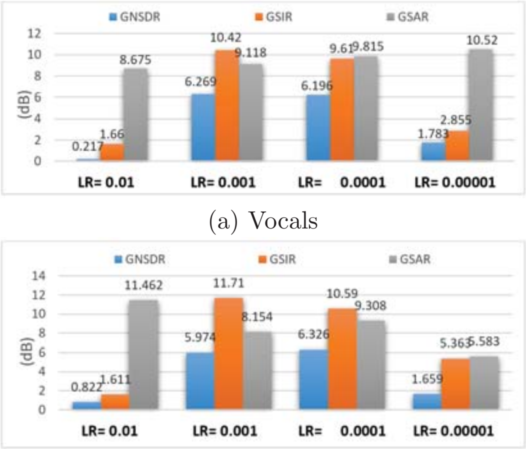


Figure 3: Selection number of layers for the hybrid model.

Phase two: The effect of learning rate. Fig. 4 reports the results by varying the value of learning rate within the range 0.01 – 0.00001 makes almost not a big difference in GSAR value, while it affects the GNSDR and GSIR values. It has been observed that models with a learning rate of 0.001 provided higher GNSDRs, GSIR but lower GSARs, compared to other cases. Thus, the learning rate value is fixed at 0.001 in the following experiments.

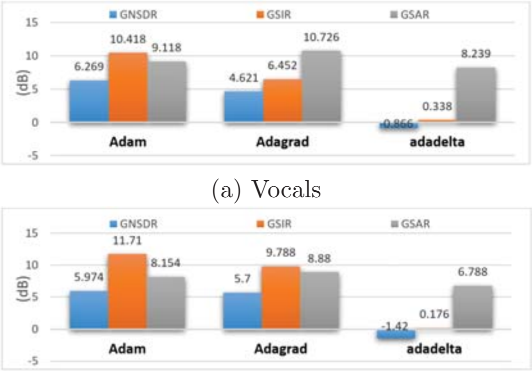


(a) Vocals

(b) Accompaniment

Figure 4: The separation effect of using different learning rate values (a) The comparison of the estimated vocals. (b) The comparison of estimated accompaniment.

Phase three: The effect of optimizer. Optimization algorithms are responsible for reducing the square error between the predicted and original source; furthermore, it provides the most accurate results possible. Fig. 5 reports the results using the model with different optimization algorithms (Adam, Adagrad, and Adadelta, respectively). As can be seen, Adam produces better results when compared to the other scenarios. As a result, we fix the optimizer Adam in the following experiments.

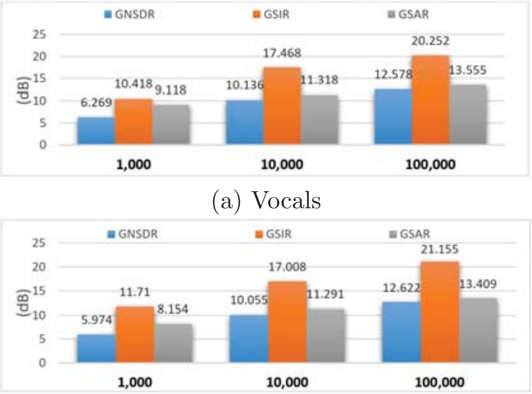


(a) Vocals

(b) Accompaniment

Figure 5: The separation effect of using different optimization algorithms (a) The comparison of the estimated vocals. (b) The comparison of estimated accompaniment.

Phase four: The effect of the increasing number of epochs. Fig 6 shows the difference when using a different number of epochs to train the model. We explore the cases with (1000, 10,000, 100,000 epochs). As can be seen, the separation performance improves with increasing the number of epochs.



(a) Vocals

(b) Accompaniment

Figure 6: The separation effect of using different number of epochs (a) The comparison of the estimated vocals. (b) The comparison of estimated accompaniment.

Finally, the optimal results in this paper are compared to other algorithms. Fig. 7 and Table 2 list the results with the supervised and unsupervised settings. From Fig. 7 and Table 2 we can observe that we get 4.81 dB gain in GNSDR, 7.28 dB gain in GSIR, and 3.39 GSAR dB gain compared to previous research finding. Fig. 8 shows an example of the separation results in time domain. From Fig. 8 we can realize that the estimated sources from the separation process are approximately the same as the clean sources which mean they are clearly estimated. Fig. 9 shows an example of the separation results in spectrogram domain. From Fig. 9 we can realize that the spectrogram of the estimated sources from the separation process are approximately the same as the spectrogram of the clean sources which mean they are clearly estimated.

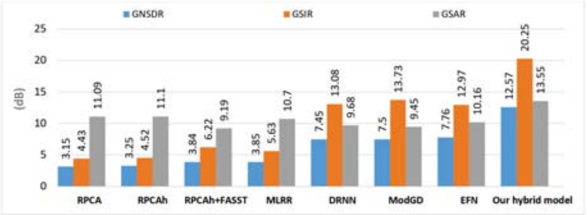


Figure 7: The separation effects on vocal between our proposed model and different approaches.

Table 2: The comparisons of the singing voices that have been separated under different approaches.

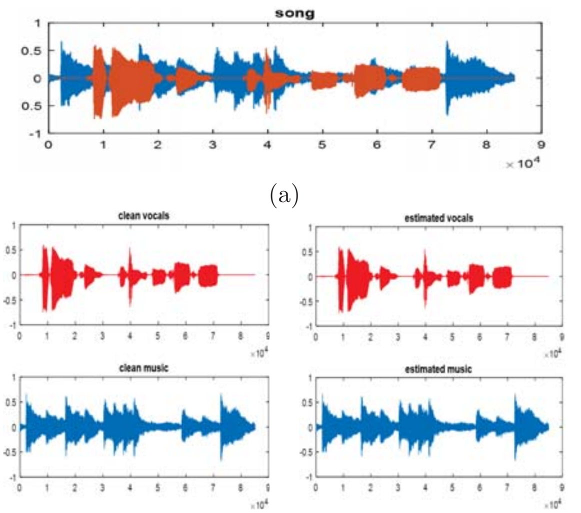
Model	Year	GNSDR (dB)	GSIR (dB)	GSAR (dB)
RPCA [7]	2012	3.15	4.43	11.09
RPCAh [28]	2012	3.25	4.52	11.1
RPCAh+FASST [28]	2012	3.84	6.22	9.19
MLRR [29]	2013	3.85	5.63	10.7
DRNN [9]	2015	7.45	13.08	9.68
ModGD [16]	2016	7.5	13.73	9.45
EFN [30]	2019	7.76	12.97	10.16
Our hybrid model	2021	**12.57**	**20.25**	**13.55**

5 Conclusion

In this study, a hybrid deep learning model has been proposed that was based on the convolutional neural network, dense neural network and recurrent neural network for the separation of the singing voice (vocals) from the monaural recordings. A trial and error method has been proposed to optimize the number of layers in each phase of the hybrid model. Furthermore, results have been enhanced by optimizing the soft mask function with the proposed model. In addition, different parameters have been tackled, such as; learning rate, the number of epochs, and different optimization algorithms. In which the architecture of the proposed system reduces the time required for training as we extract the features in CNN before feeding it to RNN for separation. Overall, our proposed model achieved (4.81) dB GNSDR gain, (7.28) dB GSIR gain, and (3.39) dB GSAR gain in comparison with baseline RNN and other separation algorithms.

References

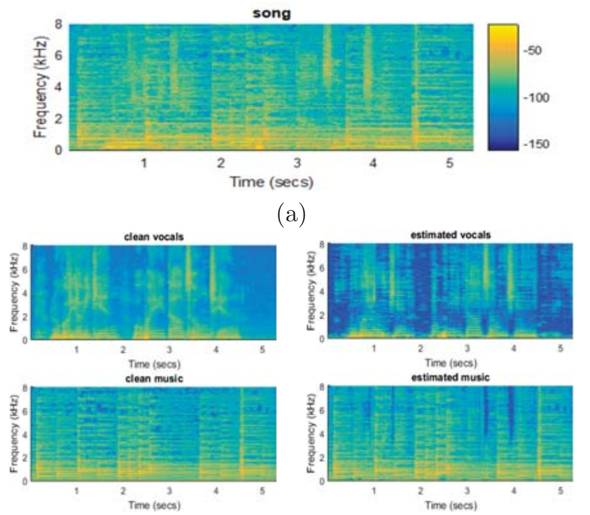
[1] Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 22*, 10 (2014), 1533–1545.



(a)

(b) (c)

Figure 8: (a) The mixture signal. (b) The clean sources. (c) The estimated sources.



(a)

(b) (c)

Figure 9: (a) The mixture signal. (b) The clean sources. (c) The estimated sources.

[2] Al-Tmeme, A., Woo, W. L., Dlay, S. S., and Gao, B. Underdetermined convolutive source separation using gem-mu with variational approximated optimum model order nmf2d. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 25*, 1 (2017), 35–49.

[3] Al-Tmeme, A., Woo, W. L., Dlay, S. S., and Gao, B. Single channel informed signal separation using artificial-stereophonic mixtures and exemplar-guided matrix factor deconvolution. *Int. J. Adapt. Control Signal Process. 32*, 9 (sep 2018), 1259–1281.

[4] Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res. 12*, null (jul 2011), 2121–2159.

[5] Hermans, M., and Schrauwen, B. Training and analyzing deep recurrent neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1* (Red Hook, NY, USA, 2013), NIPS'13, Curran Associates Inc., p. 190–198.

[6] Hsu, C., and Jang, J. R. On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 18*, 2 (2010), 310–319.

[7] Huang, P., Chen, S., Smaragdis, P., and Hasegawa-Johnson, M. Singing-voice separation from monaural recordings using robust principal component analysis. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2012), pp. 57–60.

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

[9] Huang, P., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 23*, 12 (2015), 2136–2147.

[10] Huang, P.-S., Kim, M., Hasegawa-Johnson, M. A., and Smaragdis, P. Singing-voice separation from monaural recordings using deep recurrent neural networks. In *ISMIR* (2014), pp. 477–482.

[11] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2015).

[12] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM 60*, 6 (may 2017), 84–90.

[13] Luo, Y., and Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 27*, 8 (2019), 1256–1266.

[14] Nugraha, A. A., Liutkus, A., and Vincent, E. Multichannel music separation with deep neural networks. In *2016 24th European Signal Processing Conference (EUSIPCO)* (2016), pp. 1748–1752.

[15] Parveen, S., and Green, P. Speech enhancement with missing data techniques using recurrent neural networks. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (2004), pp. 733–736.

[16] Sebastian, J., and Murthy, H. A. Group delay based music source separation using deep recurrent neural networks. In *2016 International*

Conference on Signal Processing and Communications (SPCOM) (2016), pp. 1–5.

[17] Shi, Z., Lin, H., Liu, L., Liu, R., Hayakawa, S., and Han, J. Furcax: End-to-end monaural speech separation based on deep gated (de)convolutional neural networks with adversarial example training. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), pp. 6985–6989.

[18] Simpson, A. J. Probabilistic binary-mask cocktail-party source separation in a convolutional deep neural network. *ArXiv abs/1503.06962* (2015).

[19] Simpson, A. J. R., Roma, G., and Plumbley, M. D. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *Latent Variable Analysis and Signal Separation* (Cham, 2015), E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds., Springer International Publishing, pp. 429–436.

[20] Slizovskaia, O., Haro, G., and Gómez, E. Conditioned source separation for musical instrument performances. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 29* (2021), 2083–2095.

[21] Sun, Y., Wang, W., Chambers, J., and Naqvi, S. M. Two-stage monaural source separation in reverberant room environments using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 27*, 1 (2019), 125–139.

[22] Tmeme, A. A., lok Woo, W., Dlay, S. S., and Gao, B. Underdetermined reverberant acoustic source separation using weighted full-rank nonnegative tensor models. *The Journal of the Acoustical Society of America 138*, 6 (2015), 3411–3426.

[23] Uhlich, S., Giron, F., and Mitsufuji, Y. Deep neural network based instrument extraction from music. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), pp. 2135–2139.

[24] Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., and Mitsufuji, Y. Improving music source separation based on deep neural networks through data augmentation and network blending. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), pp. 261–265.

[25] Wang, D. Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in Amplification 12*, 4 (2008), 332–353. PMID: 18974204.

[26] Wang, Y., Narayanan, A., and Wang, D. On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 22*, 12 (2014), 1849–1858.

[27] Woo, W., Dlay, S., Al-Tmeme, A., and Gao, B. Reverberant signal separation using optimized complex sparse nonnegative tensor deconvolution on spectral covariance matrix. *Digital Signal Processing 83* (2018), 9–23.

[28] Yang, Y.-H. On sparse and low-rank matrix decomposition for singing voice separation. In *Proceedings of the 20th ACM International Conference on Multimedia* (New York, NY, USA, 2012), MM '12, Association for Computing Machinery, p. 757–760.

[29] Yang, Y.-H. Low-rank representation of both singing voice and music accompaniment via learned dictionaries. In *ISMIR* (2013), pp. 427–432.

[30] Yuan, W., He, B., Wang, S., Wang, J., and Unoki, M. Enhanced feature network for monaural singing voice separation. *Speech Communication 106* (2019), 1–6.

[31] Zeiler, M. D. Adadelta: An adaptive learning rate method. *ArXiv abs/1212.5701* (2012).