Monaural Singing Voice Separation Using Fusion-Net with Time-Frequency Masking

Feng Li*, Kaizhi Qian†, Mark Hasegawa-Johnson†, Masato Akagi*
*Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
†Beckman Institute, University of Illinois at Urbana-Champaign, USA
*{lifeng, akagi}@jaist.ac.jp; †{kqian3, jhasegawa}@illinois.edu

***Abstract*—Monaural singing voice separation has received much attention in recent years. In this paper, we propose a novel neural network architecture for monaural singing voice separation, Fusion-Net, which is combining U-Net with the residual convolutional neural network to develop a much deeper neural network architecture with summation-based skip connections. In addition, we apply time-frequency masking to improve the separation results. Finally, we integrate the phase spectra with magnitude spectra as the post-processing to optimize the separated singing voice from the mixture music. Experimental results demonstrate that the proposed method can achieve better separation performance than the previous U-Net architecture on the ccMixter database.**

I. Introduction

Over the past several decades, monaural singing voice separation has become a hot research topic in the context of audio signal processing [1] [2]. The target is to separate an individual singing voice from the musical mixture. It has a wide range of applications such as music information retrieval (MIR) [3], singer identification [4], karaoke application [5], and leading instrument detection [6]. Various approaches have been introduced so far such as Non-negative Matrix Factorization (NMF) [7] [8] [9], kernel additive modeling (KAM) [10], Repeating Pattern Extraction Technique (REPET) [11], Robust Principal Component Analysis (RPCA) [12], and the combinations or deformations of those separation approaches [13] [14]. However, the separation results of state-of-the-art methods are still far behind human hearing capability. The existing problems of singing voice separation are still facing severe challenging [15] [16]. To obtain better separation results, Yang [17] proposed a new algorithm called multiple low-rank representation (MLRR) to decompose a magnitude spectrogram into two low-rank matrices, which is advantageous in that potentially more training database can be harvested to improve the separation result. Deep learning [18] [19] based monaural singing voice separation has been proven the significant improvement in the separation performance than the previous methods. Additionally, Convolutional Network Network (CNN) architecture has been successful in audio source separation, especially in singing voice separation [20] [21] [22]. Chandna et al. [20] utilized the convolutional filters specifically for audio database and allowed a significant gain in processing time over a simple multi-layer perception, in the fully connected layer, dimensional reduction allows the model to learn a more compact representation of the input data from which the source can be separated. Takahashi et al. [21] extended DenseNet to tackle the music source separation with the proposed MDenseNet architecture. In addition, he [22] proposed another MMDenseLSTM framework for audio source separation, which is a variant of CNN architecture. It integrates long short-term memory (LSTM) in multiple scales with skip connection to efficiently model long-term structures within an audio context.

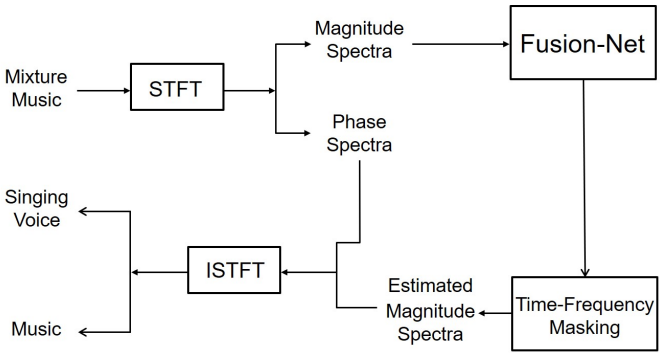

Fig. 1. Block diagram of monaural singing voice separation system

With the development of neural network architecture, the separation performance based on CNN framework has also obtained a better improvement, especially for U-Net [23] architecture in singing voice separation task. It separates singing voice from the mixture music database by using down-sampling and up-sampling frameworks on the magnitude spectrogram. The experiment results show this approach can bring clear improvements over state-of-the-art approaches. The benefits of low-level skip connections are demonstrated in comparison to plain convolutional encoder-decoders. However, there is still existing plenty of room for improving the separation performance by developing a much deeper neural network than U-Net architecture.

Therefore, this work was inspired by U-Net architecture for singing voice separation and utilized the FusionNet [24] architecture, which was originally proposed to solve image separation in connectomics. We utilized the summation-based skip connections to develop a much deeper network architecture, which can bring significant improvements in separation performance in singing voice separation task. In addition, we applied time-frequency masking to improve the separation

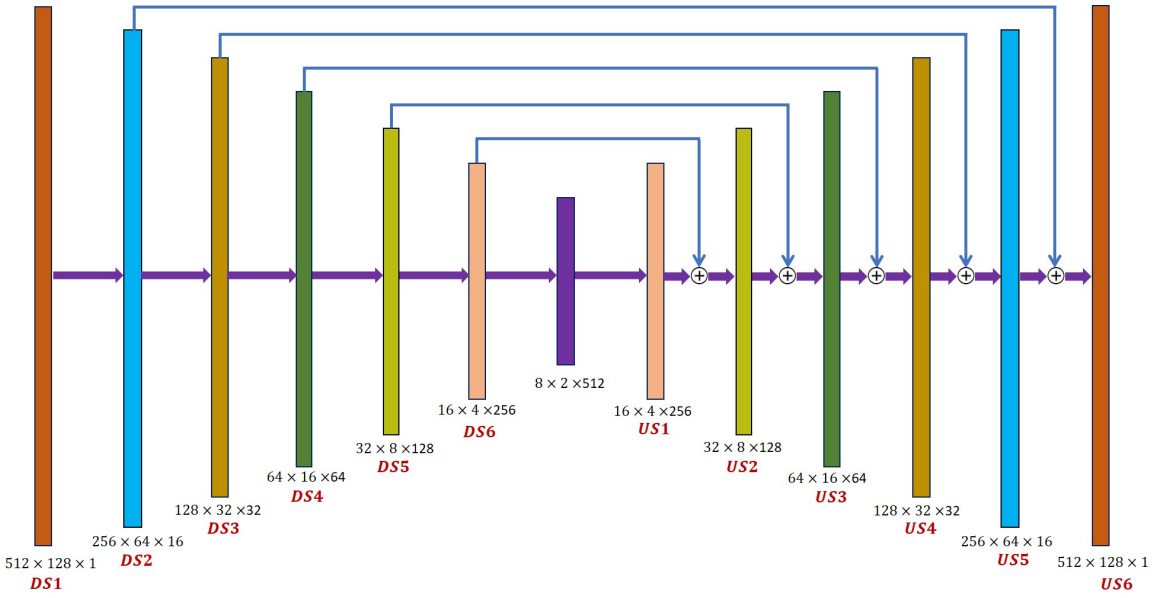

Fig. 2. Proposed Fusion-Net architecture

results. Finally, we integrated the phase spectra feature with magnitude spectra feature as the post-processing to optimize the separated singing voice by using Fusion-Net architecture from the mixture music database.

The block diagram of our proposed monaural singing voice separation system can be seen in Fig. 1. For each of mixture music on the test audio database, firstly, we applied the short-time Fourier transform (STFT) to obtain the magnitude spectra and phase spectra. Then, we explored the time-frequency masking to further improve the separation results by using the introduced Fusion-Net architecture on the magnitude spectrogram. Finally, we utilized the inverse STFT (ISTFT) between the phase spectra and estimated magnitude spectra to obtain the singing voice from the mixture music database.

The contributions of this paper can be summarized as follows:

- A deep fully residual convolutional neural network was introduced for monaural singing voice separation by combining U-Net with the residual convolutional neural network to develop a much deeper architecture with summation-based skip connections.
- Using time-frequency masking for improving the separated singing voice.
- integrating the phase spectra with magnitude spectra as the post-processing. And comparing with the separation results by proposed Fusion-Net and U-Net architectures on the ccMixter database.

The remainder of this paper is structured as follows: In Section II, we introduce the proposed method. Experiments are conducted in Section III, and finally draw conclusions in Section IV.

II. Proposed Method

In this section, we first introduce the proposed network architecture and then explain its application to singing voice separation with time-frequency masking.

A. Network Architecture

We explore a deep fully residual convolutional neural network to develop a much deeper neural network than U-Net architecture. Similar to Quan et al. [24] proposed the FusionNet architecture, which is originally used for image segmentation in connectomics. In this work, we propose the Fusion-Net architecture for singing voice separation, which is based on encoder (e.g., down-sampling: *DS*) and decoder (e.g., up-sampling: *US*). The framework of this architecture and each of the blocks can be seen in Fig. 2, which presents the detail realized process of the proposed Fusion-Net architecture.

For the fair comparison, the implementation of the proposed Fusion-Net architecture is similar to [23]. Each of blocks in *DS* consists of a strided 2D convolution, kernel size 5×5, leaky rectified linear units (ReLU) with leakiness 0.2, and batch normalization. During the process of *DS*, it contains conv2d with stride 1, kernel size 5×5, residual layers, and max-pooling. Meanwhile, each of block in *US* consists of strided deconvolution with stride 2, kernel size 5×5, and batch normalization. During the process of *US*, it contains a 2D deconvolutional layer with stride 1 and kernel size 5×5, and inverse residual layer. This model is trained by using the ADAM optimizer [25].

In addition, the detail parameters about output and input sizes in Fusion-Net architecture are described in Table I. The important differences between U-Net and Fusion-Net are skip-connection. U-Net adopts the concatenation of feature maps via only the skip connection, while Fusion-Net uses a fully residual network with summation-based skip connection in the

TABLE I
ARCHITECTURE OF THE FUSION-NET

Funion-Net architecture					
Down-sampling (*DS*)			Up-sampling (*US*)		
Blocks	Input size	Output size	Blocks	Input size	Output size
DS1	512 × 128 × 1	256 × 64 × 16	*US1*	16 × 4 × 256	32 × 8 × 128
DS2	256 × 64 × 16	128 × 32 × 32	*US2*	32 × 8 × 128	64 × 16 × 64
DS3	128 × 32 × 32	64 × 16 × 64	*US3*	64 × 16 × 64	128 × 32 × 32
DS4	64 × 16 × 64	32 × 8 × 128	*US4*	128 × 32 × 32	256 × 64 × 16
DS5	32 × 8 × 128	16 × 4 × 256	*US5*	256 × 64 × 16	512 × 128 × 1
DS6	16 × 4 × 256	8 × 2 × 512	*US6*	512 × 128 × 1	512 × 128 × 1

TABLE II
NETWORK IN EACH OF BLOCKS.

DS(1-6)		***US(1-6)***	
Layers		Layers	
conv2d		deconv2d	
Max-Pooling		deconv2d	
Residual Layer	conv2d	Inverse Residual Layer	deconv2d
	conv2d		deconv2d
	conv2d		deconv2d
Max-Pooling		deconv2d	
conv2d			
Max-Pooling			

deeper network architecture. For example, *DS6* and *US1* are summed up as the input feature in *DS2*.

The parameters of each of the blocks (*DS* and *US*) can be seen in Table II. The *DS* is the process of down-sampling, which is from *DS1* to *DS6*. Meanwhile, the *US* is the process of up-sampling, which is from *US1* to *US6*. The residual layer and inverse residual layer are included in the processes of down-sampling and up-sampling, respectively. The right is process of down-sampling section and the corresponding left is process of up-sampling section in the Fusion-Net architecture. Fig. 3 shows an example of the process of down-sampling (e.g., *DS6*) and up-sampling (e.g., *US1*) in the Fusion-Net architecture.

In this work, we adopt the training model with the predict value of the network $\hat{y_i}$ and the target value y_i, the mean values of loss function in the Fusion-Net architecture can be defined as

$$L = ||\hat{y_1} - y_1|| + ||\hat{y_2} - y_2||; \tag{1}$$

where $\hat{y_1}$ and $\hat{y_2}$ are the predict values of singing voice and music, respectively.

B. Time-Frequency Masking

In order to improve the separation performance, after separated by using Fusion-Net architecture, we further apply soft time-frquency masking estimation to improve the separation results. We define it as follows

$$\begin{aligned} \hat{y_1} &= \frac{\hat{y_1}}{\hat{y_1} + \hat{y_2}} \odot X \\ \hat{y_2} &= \frac{\hat{y_2}}{\hat{y_1} + \hat{y_2}} \odot X \end{aligned} \tag{2}$$

where the operator $\odot$ indicates the element-wise multiply (Hadamard product), X is the value of magnitude spectra, $\hat{y_1}$ and $\hat{y_2}$ are the corresponding predict values of singing voice and music, respectively.

In order to recover the singing voice and accompaniment by using ISTFT, we combine phase spectra [26] with estimated magnitude spectra Y. The phase spectra P can be defined as

$$P = angle(X); \tag{3}$$

Therefore, the recovered spectrogram $\widetilde{X}$ by combining phase spectra and estimated magnitude spectra in the complex coordinate can be obtained as

$$\widetilde{X} = Y \odot cos(P) + i(Y \odot sin(P)), \tag{4}$$

where the operator $\odot$ indicates the element-wise multiply.

III. EXPERIMENTAL EVALUATION

In this section, we evaluate the proposed Fusion-Net architecture on the ccMixter [13][1] database.

A. Experiment Setups

To confirm the effectiveness of separation performance with the proposed Fusion-Net architecture, we evaluated it on the ccMixter database, which consists of 50 tracks stereo music songs. Each of tracks is ranging from 1'17" to 7'36". And they were sampled at 44.1 kHz. In this experiment, we mainly research on monaural singing voice separation, which is generally even more difficult than multichannel due to the availability of only one channel. So, the two-channel stereo mixture experiment databases were downmixed into a single channel.

In this work, we used the experiment database as the ratio of 3:1:1 on the training, validation and testing database, in other words, 30 tracks for training, 10 tracks for validation, and left 10 tracks for testing, respectively. The experiment environments were run by using TensorFlow framework[2] and NVIDIA GeForce GTX 1080Ti with i7-6700K CPU@4.00 GHz.

We assessed its separation performance in terms of source-to-distortion ratio (SDR), source-to-interference ratio (SIR), source-to-artifact ratio (SAR), and normalized SDR (NSDR)

[1]https://members.loria.fr/ALiutkus/kam/
[2]https://www.tensorflow.org/

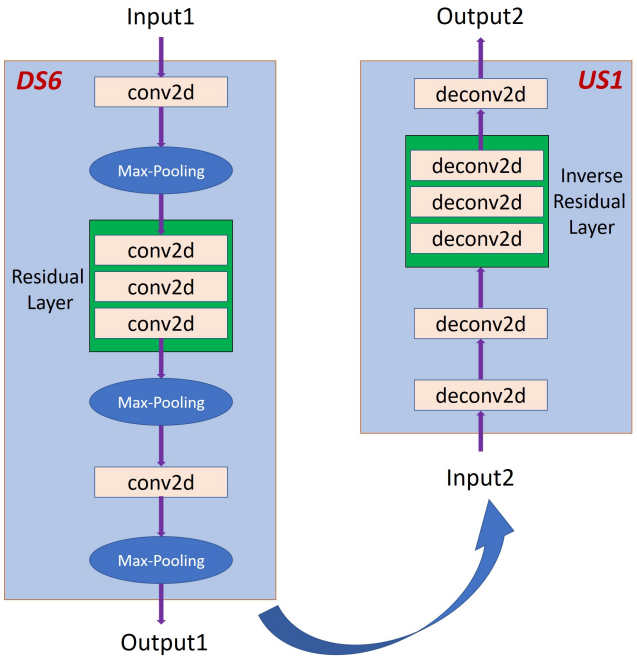

Fig. 3. Process of down-sampling (e.g., *DS6*) and up-sampling (e.g., *US1*) in Fusion-Net architecture.

by using the BSS-EVAL 3.0 metrics [27][3]. The estimated signal $\hat{S}$(t) is defined as

$$\hat{S}(t) = S_{target}(t) + S_{interf}(t) + S_{artif}(t), \quad (5)$$

where $S_{target}(t)$ is the allowable deformation of the target sound, $S_{interf}(t)$ is the allowable deformation of the sources that account for the interferences of the undesired sources, and $S_{artif}(t)$ is an artifact term that may correspond to the artifact of the separation method. The formulas for SDR, SIR, SAR, and NSDR are defined as

$$SDR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t \left(S_{interf}(t) + S_{artif}(t)\right)^2}, \quad (6)$$

$$SIR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t S_{interf}(t)^2}, \quad (7)$$

$$SAR = 10\log_{10}\frac{\sum_t \left(S_{target}(t) + e_{interf}(t)\right)^2}{\sum_t e_{artif}(t)^2}, \quad (8)$$

and

$$NSDR(\hat{v}, v, x) = SDR(\hat{v}, v) - SDR(x, v), \quad (9)$$

where $\hat{v}$ is the separated voice part, v is the original singing voice signal, and x is the original mixture value. The NSDR is used to estimate the overall improvement in SDR between x and $\hat{v}$.

Higher values of SDR, SIR, SAR, and NSDR mean that the method exhibits better separation performance in terms of the singing voice separation tasks. More specifically, the value of SDR indicates the overall quality of the separated target sound signals. And the value of SIR reflects the suppression of the interfering source, while the value of SAR represents the absence of artificial distortion. All metrics are calculated in dB.

[3]http://bass-db.gforge.inria.fr/bss_eval/

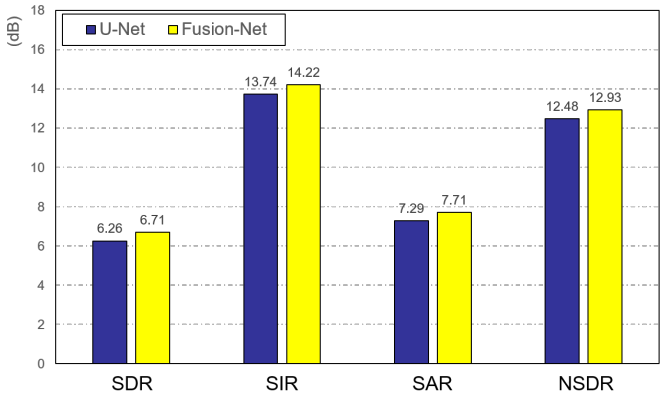

Fig. 4. Comparison of monaural singing voice separation results on the **ccMixter** database by using U-Net and Fusion-Net architectures in all metrics of SDR, SIR, SAR, and NSDR, respectively.

B. Experiment Results

Fig. 4 shows the comparison of monaural singing voice separation results on the ccMixter database between U-Net and Fusion-Net on the separation metrics of SDR, SIR, SAR, and NSDR, respectively. The experimental evaluation results clearly reveal that the proposed Fusion-Net architecture has a better separation performance than U-Net architecture for singing voice separation on the ccMixter database in all evaluation metrics.

IV. Conclusions

In this paper, we have proposed a novel monaural singing voice separation approach by exploring the proposed Fusion-Net architecture with time-frequency masking under the phase spectra and magnitude spectra. Experimental results on the ccMixter database indicate that the proposed Fusion-Net architecture outperforms U-Net architecture. For future work, since F0 estimation and melody extraction are very crucial for separating singing voice from the mixture music signal, therefore, we will unify among of them to improve the separation performance from the more complex mixture audio database.

Acknowledgment

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan Scholarship and the China Scholarship Council (CSC) of China Scholarship.

REFERENCES

[1] Z. Rafii, A. Liutkus, F.R. Stöter, S.I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 8, pp. 1307-1335, 2018.

[2] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.R. Stöter, "Musical source separation: An introduction," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp.31-40, 2019.

[3] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: current directions and future challenges," in: *proceedings of the IEEE*, vol. 96, no. 4, pp. 668-696, 2008.

[4] M. N. Chinthaka, C.S. Xu, and Y. Wang, "Singer identification based on vocal and instrumental models," in: *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, pp. 375-378, 2004.

[5] A. J. R. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: extracting vocals from musical mixtures using a convolutional deep neural network," in: *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Springer, Cham, pp. 429-436, 2015.

[6] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1180–1191, 2011.

[7] T. O. 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] M.N. Schmidt and M. Mϕrup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in: *Proceedings of Independent Component Analysis and Blind Signal Separation (ICA)*, pp. 700-707, 2006.

[9] A. Chanrungutai and C. A. Ratanamahatana, "Singing voice separation for mono-channel music using non-negative matrix factorization," in: *Proceedings of International Conference on Advanced Technologies for Communications*, pp. 243-246, 2008.

[10] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, 2014.

[11] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): a simple method for music/voice separation," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 1, pp. 73-84, 2013.

[12] P. S. Huang, S. D. Chen, P. Smaragdis and M. H. Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in: *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 57-60, 2012.

[13] A. Liutkus, D. Fitzgerald, and Z. Rafii, "Scalable audio separation with light kernel additive modelling," in: *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 76–80, 2015.

[14] F. Li and M. Akagi, "Blind monaural singing voice separation using rank-1 constraint robust principal component analysis and vocal activity detection," *Neurocomputing*, vol. 350, pp. 44-52, 2019.

[15] A. Liutkus, F. R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono and J. Fontecave, "The 2016 signal separation evaluation campaign," in: *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Springer, Cham, pp. 323-332, 2017.

[16] F. R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in: *International Conference on Latent Variable Analysis and Signal Separation*, Springer, pp. 293-305, 2018.

[17] Y. H Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in: *Proceedings of 18th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 427-432, 2013.

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

[19] J.R. Hershey, Z. Chen, J.L. Roux and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in: *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31-35, 2016.

[20] P. Chandna, M. Miron, J. Janer, and E. Gomez, "Monoaural audio source separation using deep convolutional neural networks," in: *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2017.

[21] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band DenseNets for audio source separation," in: *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 21–25, 2017.

[22] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," *arXiv preprint arXiv*:1805.02410, 2018.

[23] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in: *Proceedings of 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, pp. 745-751, 2017.

[24] T. M. Quan, D. G. Hilderbrand, and W.-K. Jeong, "Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics," *arXiv preprint arXiv*:1612.05360, 2016.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv*:1412.6980, 2014.

[26] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107-115, 2014.

[27] 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.