International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 2
ISSN: 2088-5334

Music Source Separation Using ASPP Based on Coupled U-Net Model

Suwon Yang[a], Daewon Lee[b,*]

[a] Department of Computer Science and Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 02841, Korea
[b] Department of Computer Engineering, Seokyeong University, 124 Seogyeong-ro Seongbuk-gu, Seoul, 02173, Korea
Corresponding author: *daelee@skuniv.ac.kr

***Abstract*—Noise has established itself as one of the factors that interfere with modern human life, and various noise canceling techniques have been studied to prevent noise. While the old era's noise-canceling technique focused on the physical soundproofing technique, multiple studies have been conducted on the active noise canceling technique that removes only the activated noise in the current era. Active noise canceling (ANC) or digital noise-canceling technology is based on the sound source separation method. This leads to sound source separation technology, which refers to the technology to separate individual sound signals from mixture sounds. Most of the source separation technologies focus on improving speech, not noise reduction. This technology makes it possible to obtain desired sound information more accurately and further improves noise-canceling technology by eliminating unwanted sound information. To provide deeper capability and more enhanced sound separation than the existing structure, we are focused on coupled U-Net model and Atrous spatial pyramid pooling technique (ASPP). This paper presents the music source separation method that combined Coupled U-Net structure with Atrous spatial pyramid pooling technique. To prove the proposed source separation method, we compared GNSDR, GSIR, and GSAR using MIR-1K, a data set that can evaluate the performance of the music source separation. Performance results show that the proposed source separation method overcame other methods' disadvantages and strengthened the feature map.**

***Keywords*—Active noise canceling; sound source separation; music source separation; U-net structure; coupled U-Net structure; Atrous Spatial Pyramid Pooling.**

Manuscript received 6 Aug. 2020; revised 29 Dec. 2020; accepted 24 Feb. 2021. Date of publication 30 Apr. 2021.

I. INTRODUCTION

Recently, noise-canceling technology removes noise from the surrounding area and is applied and needed in everyday life. Especially, the technology is used in various fields such as voice recognition technology, medical fields such as hearing aids, and voice device such as earphone headsets. Most of the source separation technologies focus on improving speech, not noise reduction. This leads to sound separation technology that separates individual sound signals from mixture sounds. This technology makes it possible to obtain desired sound information more accurately and further improves noise-canceling technology by eliminating unwanted sound information.

In this paper, we propose a Coupled U-Net [1] based structure. The proposed structure is capable of deeper and more enhanced sound separation than the existing FCN(Fully Connected Network)based structure using the atrous spatial pyramid pooling. Performance evaluations such as GNSDR, GSIR, and GSAR were compared using MIR-1K, a data set that can evaluate the performance of the music source separation.

II. MATERIALS AND METHOD

A. Analytical Model

Table 1 shows a feature comparison of a representative analysis model based on the U-Net structure.

TABLE I
COMPARING ANALYTICAL MODELS BASED ON U-NET STRUCTURE

Method	Resolution structure	Feature reuse	Communication field
U-Net [2]	Single	X	-
Dense U-Net [4]	Single	O	The whole the model
Wave U-Net [9]	Single	O	within the same stack
Stacked U-Net [5]	Multi	O	within the same stack
Coupled U-Net [1]	Multi	O	The whole of the model

1) U-Net: A number of noise separation studies have been conducted using machine running (deep-running), which uses the encoder-decoder structure much because sound separation technology requires sophisticated pixel-level segmentation [2]. Fig. 1 shows the U-Net structure. The typical U-Net structure problem is a relatively shallow layer, so the problem is that the semantic information can not be encoded effectively, and information about the feature can be lost.

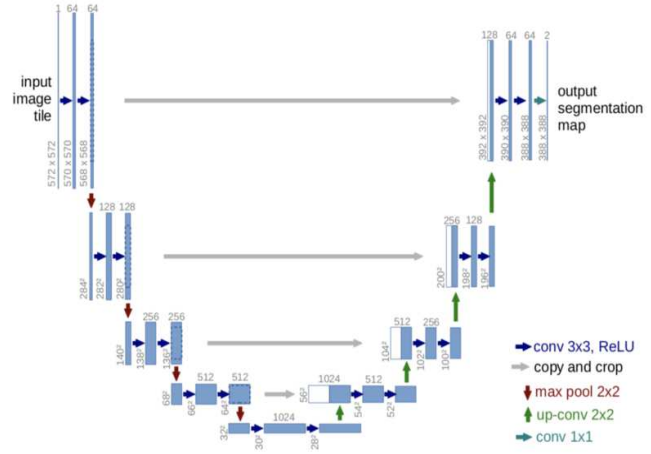

Fig. 1 U-Net structure [3]

2) Wave U-Net: Wave U-Net [9] structure is a proposed structure to overcome the difficulty of high-quality separation due to the high sampling rate of audio. The U-Net structure is adapted to a one-dimensional time zone, and the feature map is repeatedly resampled to calculate and combine features on different time scales. Fig. 2 shows the wave U-Net structure. By changing the existing convolution to a stride convolution, we avoided the artifacts problem that occurred in the existing convolution.

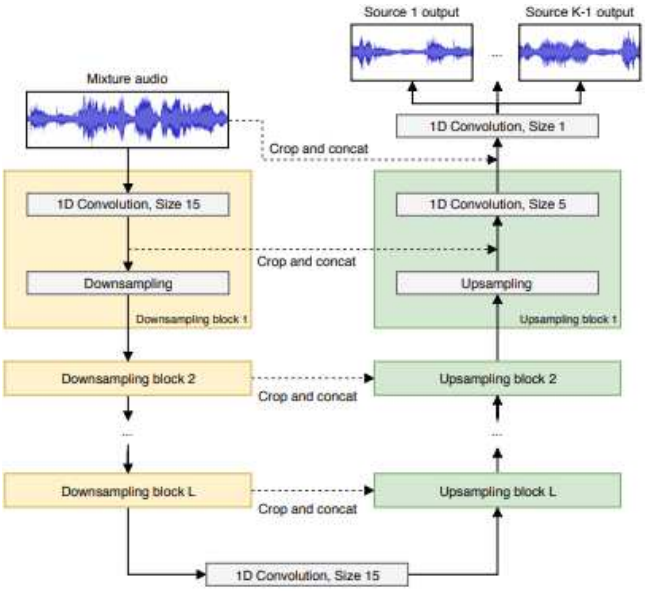

Fig. 2 Wave U-Net structure [9]

3) Dense U-Net: Dense U-Net structure continued to link the feature map of the previous layers to the next layer's input, solving the problem that the semantic information could not be encoded effectively because of the relatively shallow layer of the existing U-Net structure. It also improved parameter efficiency by reusing feature maps from previous layers [4]. Fig. 3 shows a convolution layer of dense U-Net structure. Fig. 4 shows a dense U-Net structure. However, there are problems with the multi-resolution structure and performing spatial information because that is a single resolution structure. There are problems in memory management because the previous feature maps are all saved and affected the next layers.

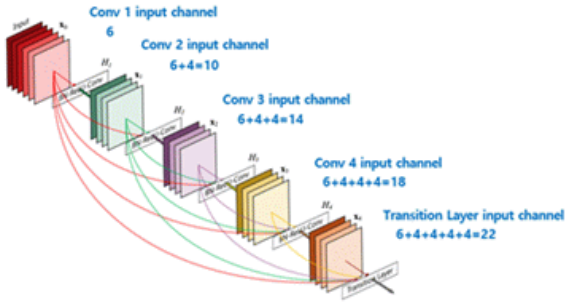

Fig. 3 Convolution layer of Dense U-Net structure [4]

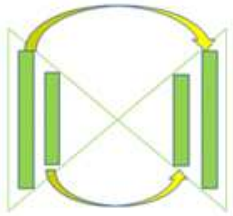

Fig. 4 Dense U-Net structure

4) Stacked U-Net: Stacked U-Net [5] structure is proposed to overcome a single resolution structure by using multiple resolution structure, and it is a structure that enables more spatial information execution. Fig. 5 shows a stacked U-Net structure. However, the above structure has a disadvantage because the influence on the information may be small because it communicates only within the same stacked structure.

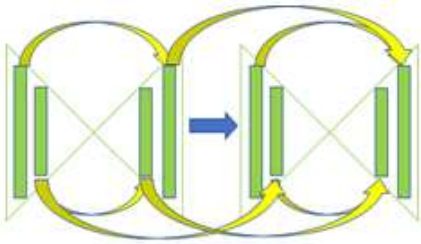

Fig. 5 Stacked U-Net structure

5) Coupled U-Net: Coupled U-Net structure is a structure that enables more spatial information execution by using multiple resolution structure like stacked U-Net structure and is a structure that reinforces a feature map by stacking without loss of resolution [1]. Fig. 6 shows a coupled U-Net structure. However, unlike the Stacked U-Net structure, it does not communicate only within the same stacked structure. To enhances the feature map by coupling up the same pair of semantic blocks in the previous stack, creates a more efficient information flow than the stacked U-Net structure. More information movement is possible, and the

efficiency of the parameters is increased by reusing the previous feature map. However, this structure also needs to save feature maps continuously, which causes memory-intensive problems.

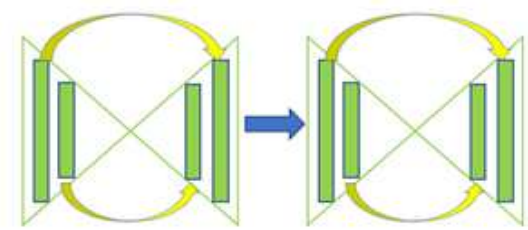

Fig. 6 Coupled U-Net structure

B. Analytical Layer

1) Atrous spatial pyramid pooling: Atrous (dilated) convolution [8] is a structure that operates by zero paddings the empty space inside the filter, unlike the FCN (Fully Connected Network). Since zero-padding processing is performed on the empty space, the operation efficiency is high, and the rate can be adjusted to increase the receptive field seen by the filter to minimize the loss of information to obtain a high-resolution output and obtain good performance. In order to better cope with this multi-scale, we can set the expansion factor for Atrous convolution and apply the result, increase the receptive field by concatenating the result, and at the same time, expect the efficiency of operation. Fig. 7 shows Atrous spatial pyramid pooling convolution.

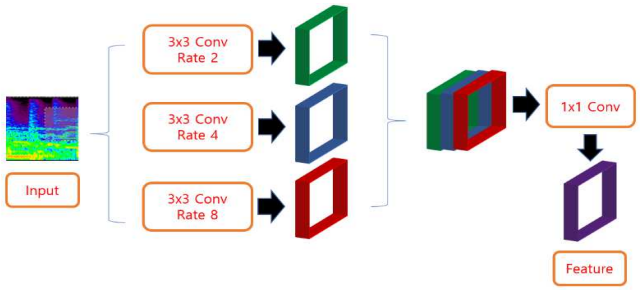

Fig. 7 Atrous spatial pyramid pooling convolution [8]

2) Depthwise separable convolution: Depthwise separable convolution is a structure in which the existing convolution filter separately processes the spatial dimension and the channel dimension separately. In this process, the number of parameters can be further reduced by sharing the necessary parameters for spatial dimension processing. Even if the two axes are separated, the final result is the result of processing both axes so that the existing convolution filter can fully replace the role. Semantic segmentation, which requires label prediction for each pixel, is more difficult. Therefore, when separable convolution is used in a situation where the CNN structure is deeper and more parameters are used to widen the receptive field, the necessary parameters of the model are needed [2]. Since the number can be greatly reduced, it can be extended to a deeper structure to improve performance or to reduce memory usage and speed up compared to the existing one. Fig. 8 shows Depthwise separable convolution.

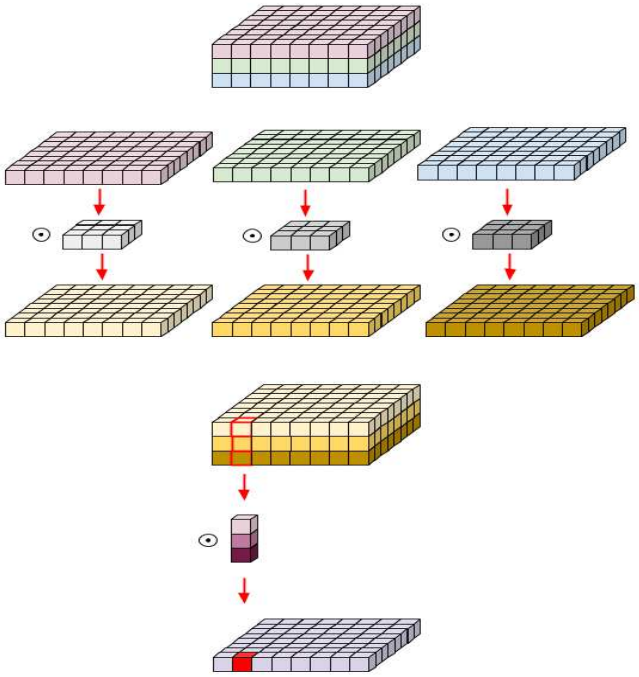

Fig. 8 Depth wise separable convolution [2]

3) Demucs: This model is based on Wave U-Net structure and consists of LSTM, convolutional encoder, and convolutional decoder. The ncoder and decoder are connected by skip U-Net [10]. Fig. 9 shows Demucs architecture with the mixture waveform. It shows input and the four sources estimates as output. Arrows represents U-Net connections.

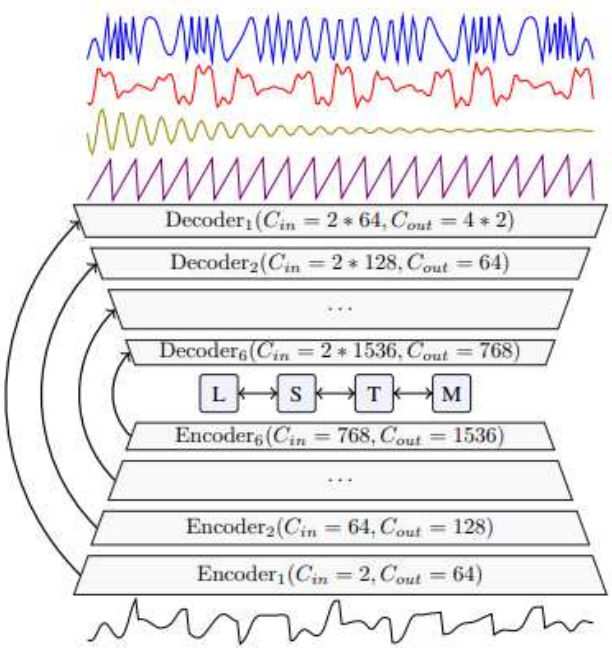

Fig. 9 Demucs architecture with the mixture waveform

Fig. 10 shows a detailed view of the layers Decoder on the top and Encoder on the bottom. Arrows represent connections to other parts of the model. Depthwise separable convolution. Batch normalization is not used because the initial experiments showed poor results for the model.

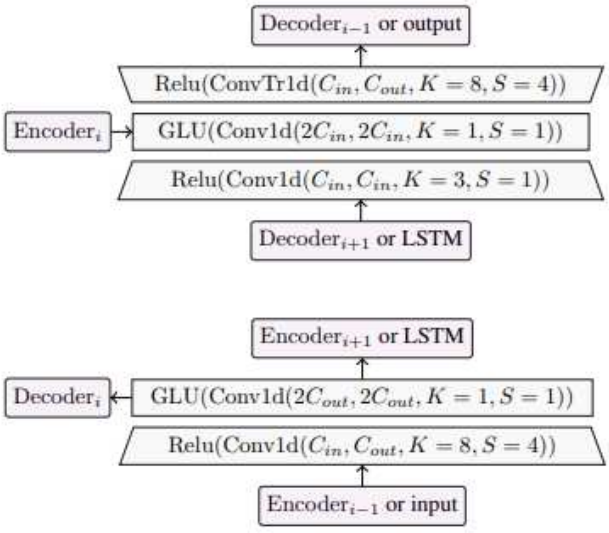

Fig. 10 Detailed representation of the encoder and decoder layers

C. *Proposed Model*

The proposed model structure is a sound separation model based on Coupled U-Net. Atrous Spatial Pyramid Pooling (ASPP) was used instead of the fully connected network (FCN) used in the previous model. Fig. 10 shows the structure of the proposed model. This has the advantage of reducing the loss of spatial dimensions, using the same number of parameters, reducing the amount of computation, and having a large field of view, thus improving the accuracy of the segmentation model, resulting in deeper and more enhanced sound separation than existing structures. Suggest this model. Depthwise separable convolution method can be used to greatly reduce the number of parameters without concatenating ASPP in a general way, so it can be extended to a deeper structure to improve performance, and memory consumption and speed improvement can be expected. Fig. 11 shows the system operation diagram.

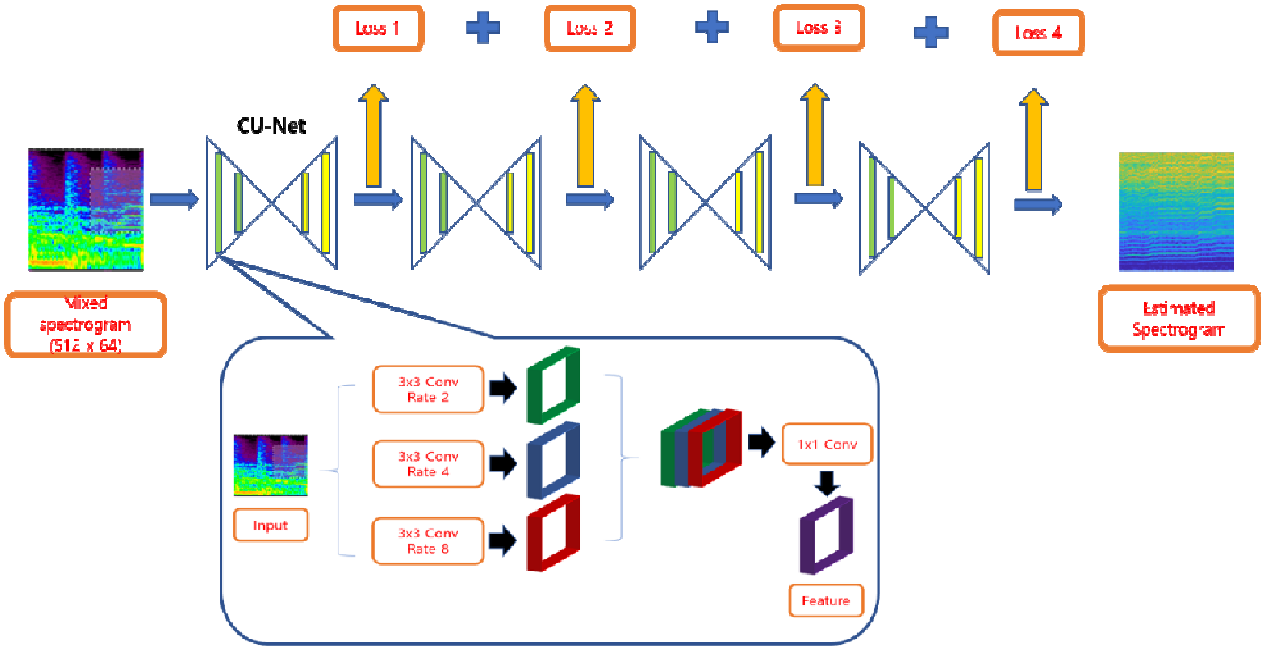

Fig. 11 Structure of proposed model

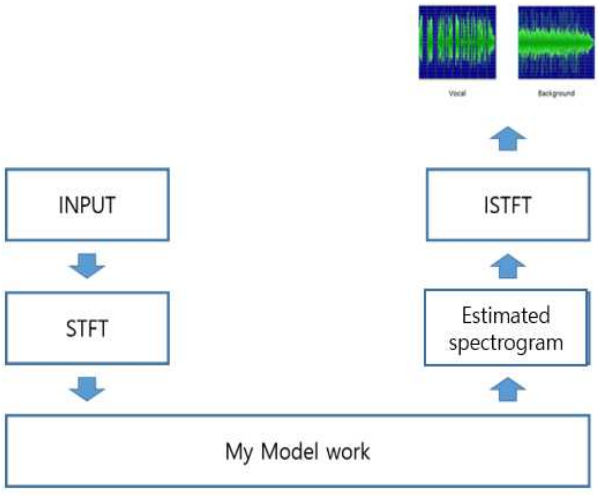

Fig. 12 System operation diagram

Using coupled U-Net structure makes it possible to extract more spatial information than a single resolution structure. Coupling semantic blocks of the same pair strengthen the feature map to make more efficient information flow, making it more efficient than the Stacked U-Net structure. A lot of information can be moved for better sound separation. Unlike the existing Coupled U-Net structure, we solved using much memory by saving and coupling only the nearest feature without saving all the features of previous layers.

III. Results and Discussion

A. *Performance Environment*

We used Ubuntu 16.04.6 and TensorFlow-GPU 1.13.1 based on python 3.6 on GeForce GTX 1050ti. MIR-1K is a dataset for singing voice separation, which exists as 1000 wav files. For the training data, 175 of one man (Adjones)

and one woman (Amy) were used as data for testing. For the performance evaluation, the values of Global Normalized-signal-to-distortion ratio (GNSDR), Global source-to-interference ratio (GSIR), and Global source-to-artifacts ratio (GSAR) were evaluated using the mir-eval toolbox.

B. Experiment result with MIR-1K

The experimental results were evaluated using the MIR-1K dataset. Table II shows the result of accompaniment, and Table III shows the result of vocal. MLRR [7] is an algorithm technique without deep learning, which shows poor performance compared to other research results using deep learning.

TABLE II
RESULT OF ACCOMPANIMENT

Method	Use Deep Learning	Model	GN SDR	GSIR	GSAR
MLRR [7]	X	-	4.19	7.80	8.22
U-Net [8]	O	U-Net	7.45	11.43	10.41
EFN [6]	O	DRNN	7.86	13.54	10.16
Stacked hourglass [5]	O	Stacked U-Net	9.88	14.24	12.36
Proposed model	O	Coupled U-Net	10.59	14.86	13.03

TABLE III
RESULT OF VOCAL

Method	Use Deep Learning	Model	GN SDR	GSIR	GSAR
MLRR [7]	X	-	3.85	5.63	10.70
U-Net [8]	O	U-Net	7.43	11.79	10.42
EFN [6]	O	DRNN	6.64	12.05	9.67
Stacked hourglass [5]	O	Stacked U-Net	10.51	16.01	12.53
Proposed model	O	Coupled U-Net	**11.19**	**16.96**	**12.91**

The U-Net structure is an encoder-decoder structure suitable for sound separation technology that requires precise pixel segmentation. However, since the deconvolutional layers are relatively shallow, there is a problem in that semantic information cannot be effectively encoded. It may result in loss of information on features. EFN is a model using DRNN structure, which avoids high spectral decomposition cost but shows similar or lower performance evaluation than other structures based on U-Net structure. Stacked hourglass is a model using a stacked U-Net structure, which enables more spatial information execution using multiple resolution structures and overcomes the disadvantages of U-Net structure by stacking and reinforcing feature maps to prevent loss of resolution. However, this has the disadvantage that the influence on the information may be small since the communication is performed only on the same stack. The model structure proposed in this paper shows the highest performance index on all surfaces. Through this, we can see that the model structure proposed in this paper overcomes the disadvantages of a stacked hourglass and strengthened the feature map.

IV. CONCLUSION

While the noise-canceling technology becomes a hot issue of interest in both the academy and the industry, a solid foundation still lacks its rapid development. Various noise-canceling devices appear, such as a headset, building, smart car system, smart home, etc. We focused on digital noise-canceling technology based on the sound source separation method. The structure using the Coupled U-Net and ASPP proposed in this paper enables more spatial information extraction than the single resolution structure and strengthens the feature map by coupling the semantic blocks of the same pair to create more efficient information flow. More information can be moved than structures, allowing for better sound separation. It is hoped that the desired sound information can be obtained more accurately than the previous technology and that the noise-canceling technology can be further improved by removing unwanted sound information.

ACKNOWLEDGMENT

This research is under Seokyeong University 2020 funding.

REFERENCES

[1] Tang, Zhiqiang, et al. "CU-net: coupled U-nets." *arXiv preprint arXiv:1808.06521* (2018).
[2] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
[3] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention.* Springer, Cham, 2015.
[4] Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[5] Park, Sungheon, et al. "Music source separation using stacked hourglass networks." *arXiv preprint arXiv:1805.08559* (2018).
[6] Yuan, Weitao, et al. "Enhanced feature network for monaural singing voice separation." *Speech Communication 106* (2019): 1-6.
[7] Yang, Yi-Hsuan. "Low-rank representation of both singing voice and music accompaniment via learned dictionaries." *ISMIR*. 2013.
[8] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. "Singing voice separation with deep u-net convolutional networks". *18th International Society for Music Information Retrieval Conferenceng,* Suzhou, China, 2017.
[9] Stoller, Daniel, Sebastian Ewert, and Simon Dixon. "Wave-u-net: A multi-scale neural network for end-to-end audio source separation." *arXiv preprint arXiv:1806.03185* (2018).
[10] Défossez, Alexandre, et al. "Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed." *arXiv preprint arXiv:1909.01174* (2019).
[11] Takahashi, Naoya, and Yuki Mitsufuji. "Multi-scale multi-band densenets for audio source separation." *2017 IEEE Workshop on*

Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2017.

[12] Stöter, Fabian-Robert, et al. "Open-unmix-a reference implementation for music source separation." (2019).

[13] Ince, Gökhan, et al. "Ego noise suppression of a robot using template subtraction." *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009.

[14] Nakajima, Hirofumi, et al. "An easily-configurable robot audition system using histogram-based recursive level estimation." *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010.

[15] Nakajima, Hirofumi, et al. "An easily-configurable robot audition system using histogram-based recursive level estimation." *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010.

[16] Luo, Yi, et al. "Deep clustering and conventional networks for music separation: Stronger together." *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[17] Yu, Dong, et al. "Permutation invariant training of deep models for speaker-independent multi-talker speech separation." *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[18] Jansson, Andreas, et al. "Singing voice separation with deep U-Net convolutional networks." (2017).

[19] Stoller, Daniel, Sebastian Ewert, and Simon Dixon. "Wave-u-net: A multi-scale neural network for end-to-end audio source separation." *arXiv preprint arXiv:1806.03185* (2018).

[20] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

[21] Tan, Ke, and DeLiang Wang. "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement." *Interspeech*. Vol. 2018. 2018.

[22] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." *Proceedings of the European conference on computer vision (ECCV)*. 2018.