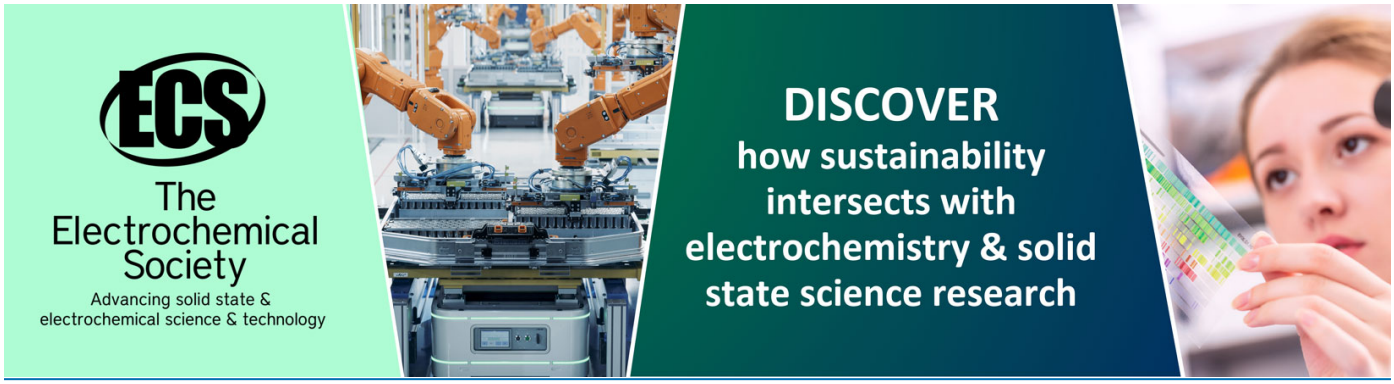
Audio source separation using supervised deep neural network



Audio source separation using supervised deep neural network

Riham J. Issa and Yusra F.Al-Irhaym

College of Computer Sciences and Mathematics
University of Mosul, Mosul, Iraq

E-mail: rihamjassim11@gmail.com

Abstract. The speech signals inserted in the computer may be mixed as a result of interference with signals from other sources. These signals may be speech signals or noise. One of the most famous examples of this problem when a group of people speaking in the same time is the "Cocktail Party". This problem produces a mixture of different speech signals, called the mixed signal. To solve this problem, the audio signals that make up the mixed-signal must be restored to their sources. This task is called separating audio sources. In this paper, supervised Deep Recurrent Neural Networks with Bi-directional Long Short Term Memory (Supervised DRNN-BLSTM) were used. To achieve a monaural source separation, we build a model to separate audio signals from a monaural mixed signal. This mixed signal consists of two different audio signals (male-female). We predict two types of time-frequency masks (Ideal Ratio Mask (IRM), and Optimal Ratio Mask (ORM). They are used to achieve the separation of the target audio sources from the mixed signal. We test the model on a dataset with (500) mixed signals. Each mixed signal three seconds in length and consists of two speaker signals (Female-Male). They are recorded in a stereo format at 8192kHz, our approach achieves Signal-to-Distortion ratio (SDR) (0.183.db), Source-to-interference Ratio (SIR) (0.198.db), and Source-to-Artifacts Ratio (SAR) (0.13.db) gain using (ORM) mask compared to the existing model using (IRM) mask.

Keywords: Monaural source separation, Ideal ratio mask (IRM), Optimal Ratio Mask (ORM), Deep Neural Network, Cocktail party problem.

1.Introduction

In the field of sound processing, separation methods are used to extract the interfering sounds of speakers in a mixture of different audio signals. The term (Source) is used on the signals forming the mixture, while the task is called Blind Source Separation (BSS)[1]. The process of separating single-channel audio sources is a special case of separating audio sources because the separation is done using a single mixed-signal. This adds another challenge as the different signals overlap in time and frequency, making the separation process more difficult [2]. Recent researches on supervised single-channel source separation as in [3,4,5], have enhanced the (BSS) techniques by merging the sources training to generate linear and nonlinear models. These models achieve an efficient source separation. The development of deep learning techniques in recent years has a significant impact on the evolution of the performance of source separation algorithms, where deep learning techniques have been used to separate different types of audio

signals, including mixed musical signals, speech signals from noise, separating interfering and synchronous speech signals[6]. The models based on Deep Neural Networks(DNN) have improved the source separation tasks because their learning capacity allows effective modeling of the interaction between the source signal and the acoustic environment in a nonlinear manner as well as the dynamic structure of speech. Discriminative features are also important[6,14].(DNNs) are usually used to predict time-frequency mask that contributes to determining the contribution of each source in the mixed signal. The Ideal mask estimation shouldn't be hard to estimate by learning machine and should obtain good separation results[10]. Separation task carried out in the time-frequency domain(TF-domain) as recovering the Short-Time Discrete Fourier Transformation (STFT) of the source signals $X_n(t,f)$ for each time frame t and frequency f , as follow[11]:

$$X(t,f) = \sum_{n=1}^{N} x[n+tL]w[n] \exp\left(-\frac{j2\pi nf}{N}\right) \quad (1)$$

The input to the network is the magnitude spectrogram of the mixed-signal assigned to training ($x_{n,f}$), and the original signals for each source ($S_{1(n,f)}$) and($S_{2(n,f)}$)[5]. After training the output prediction of the network are ($\hat{S}_{1(n,f)}$) and ($\hat{S}_{2(n.f)}$) [12].

To reconstruct estimated time-domain frames, an inverse Discrete Fourier Transform (DFT) can be used from the estimated STFT ($\hat{S}_{(n,f)}$) of each source signal. The overlap-add operation with the synthesis window v[n] is used to reconstruct the estimate ($\hat{S}_{(n,f)}$) of the target signal[11]

$$\hat{S}_{(n,f)}[n] = \frac{1}{N}\sum_{f=1}^{N} \hat{S}_s(t,f) \exp\left(\frac{j2\pi nf}{N}\right) \quad (2)$$

$$\hat{S}_{(s)}[n] = \sum_{t=1}^{T} v(n-tL)\, \hat{S}_{(s,t)}(n-tL) \quad (3)$$

2. Methodology of the Problem

The process of separating the audio source signals from single-channel mixed signal requires an estimation of (S) sources from the mixed-signal, as in equation (4).

$$x(t) = \sum_{i=1}^{S} y_i(t) \quad (4)$$

Where $y_i(t), i = 1 \ldots S$, is i^{th} of sources to be estimated, while $x(t)$ is the observed mixture[7]. For simplicity, we assume that the mixed signal is consisting of two different signals $s_1(t), s_2(t)$ as in Equation (5)

$$x(t) = s_1(t) + s_2(t) \quad (5)$$

This problem can be solved in the (STFT) Domain. Let $X(n,f)$ be the corresponding (STFT) of the mixed-signal $x(t)$, where t denotes the time domain, n represents the frames index and f is the frequency-index of the STFT domain of the signal. This problem can be formulated as follows:

$$X(n,f) = S_1(n,f) + S_2(n,f) \quad (6)$$

Where $S_1(n,f)$ and $S_2(n,f)$ are the unknown (STFTs) of the sources in the mixed signal. Given $X(n,f)$ the aim of monaural source separation is to recover one or more desired signals from the mixed signal[8]. The typical setup assumed that only (STFT) magnitude spectra is available and differences between the phase angles of the (STFT) of the sources are ignored during the separation task. This is used only when reconstruct the time domain waveforms of the sources. [9]. The magnitude spectrogram of the measured audio signal can be written as the sum of source signals magnitude spectrograms as follows:

$$|X_n| \approx |S_1(n,f)| + |S_2(n,f)| \quad (7)$$

We use the matrix form to represent the magnitude spectrograms where n and f denote the spectral frame and frequency index respectively, as follows:

$$X(n,f) \approx S_1(n,f) + S_2(n,f) \quad (8)$$

Where $S_1(n,f)$ and $S_2(n,f)$ are the unknown magnitude spectrograms of the sources that need to be estimated[8]. The magnitude spectrogram of the mixture signal $|X(n,f)|$, together with spectral features are fed into (DNN) to predict time-frequency mask for each speaker. The masks are multiplied by the mixture using (element-wise multiplication operation) to estimate the magnitude STFT of the desired speaker. The separated waveforms of the estimated speaker are resynthesized using inverse(STFT), the estimated magnitude of the speaker, and noisy phase information.[10]

3. Bi-directional Long Short Term Memory (BLSTM)

The (BLSTM) combines long short-term memory (LSTM) and bi-directional iterative recurrent neural network (BI-RNN**).** The recurrent neural network (RNN) is a special development of artificial neural networks to process continuous data. Due to the problem of gradient vanishing or explosion, the (LSTM) was created, which consists of three gates (input - forget - output). Both (RNN) and (LSTM) can obtain information from the previous context only. To solve this problem(BLSTM) was used [10]. The (BLSTM) consists of two hidden recurrent layers, receive the input string separately in two opposite directions, one in the forward direction, and the other in the backward direction. Both layers are connected to the same output layer. Thus enabling access to long contexts in two different directions [12]. In a single cell (LSTM), the hidden serial vector is calculated by calculating the output of each gate as in the following equations:

$$f_t = \sigma\left(W_f x_t + R_f h_{t-1} + b_f\right) \quad (9)$$
$$i_t = \sigma(W_i x_t + R_i h_{t-1} + b_i) \quad (10)$$
$$o_t = \sigma(W_o x_t + R_o h_{t-1} + b_o) \quad (11)$$
$$j_t = tanh\left(W_j h_t + R_j h_{t-1} + b_j\right) \quad (12)$$
$$c_t = c_{t-1} \odot f_t + j_t \odot i_t \quad (13)$$
$$h_t = \tanh(c_t) \odot o_t \quad (14)$$

Where x_t is cell's input, c_t and h_t the state and output respectively, at time t. f_t, i_t and o_t are forget gate, input gate and output gate, respectively. W, R denote the trainable weight matrices, b the bias vectors and σ the activation function of the hidden layer. In (BLSTM) the equation (14) cannot be used directly, instead, the forward hidden sequence h1 and backward hidden sequence h2 is computed as shown in equation(15) and (16) respectively[13].

$$\overrightarrow{h_1} = f(W_{x\vec{h}} x_t + W_{\vec{h}\vec{h}} \vec{h}_{t+1} + b_{\vec{h}} \quad (15)$$

$$\overleftarrow{h_2} = f(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}} \quad (16)$$

4. Mask Estimation Using Deep Neural Network(DNN)

Deep neural networks are usually used to predict a mask. This mask is used to determine the contribution of each source in the training mixed signal. The input to the network is the magnitude spectrogram of the training mixed-signal (X_{tr}), the original signals for each source (S_{tr1}) and(S_{tr2})[7]. After training, the output prediction of the network are ($\hat{S}_{tr1}$) and ($\hat{S}_{tr2}$) [13]. To smooth the spectra results, two types of masks are estimated separately by the model, the first mask is the ideal ratio mask (IRM) which is a smooth form of an ideal binary mask (IBM),(IRM) mask represented in equation (17)

$$IRM_s = \left(\frac{S_1(t,f)^2}{S_1(t,f)^2 + S_2(t,f)^2}\right)^{\beta} \qquad (17)$$

Where $S_1(t,f)^2$ and $S_2(t,f)^2$ are the energy of each speech signal in mixed-signal, in each (TF Unit). (β) is the square root value used as a tunable parameter, which is usually set to (0.5). (IRM) the mask used to conserve the speech energy of each (TF unit), assuming $S_1$and S_2 are uncorrelated[14]. The second mask is the optimal ratio mask (ORM) which can be derived by minimizing mean square error(MSE) between the clean speech signal and the target speech signal that has been estimated and can be defined by equation (18).

$$M(t,f) = \frac{|S_1(t,f)^2 + R(S_1(t,f)S_2{}^*(t,f))}{|S_1(t,f)|^2 + |S_2(t,f)|^2 + 2R(S_1(t,f)S_2{}^*(t,f))} \qquad (18)$$

Where S_1 and S_2represent the spectrum of the two speech signals in the mixed-signal at frame (t) and frequency(f). The symbol (*) indicates the conjugate operation, while R (°) represents the real component of the spectrum. The ORM differs from the IRM in the presence of the coherent part by $R(S(t,f)N^*(t,f))$, whose value is equal to zero in the IRM. The (ORM) achieves high efficiency in the separation process in the case of a high correlation between the speech signal and the noise signal. The mask ORM values range from (-∞, + ∞), which makes the estimation process more difficult. So the values of the ORM mask are determined using the hyperbolic tangent function as in equation (19).

$$ORM(t,f) = K\frac{1 - e^{-cy(t,f)}}{1 + e^{-cy(t,f)}} \qquad (19)$$

Where c=0.1 is the steepness, while K is equal to 10, it restricts the values of the ORM between (-10, + 10). Equation (18) is the basic equation of the ORM mask[14].

5. Network architecture

The (DRNN_BLSTM) architecture based on monaural speech separation was used in this paper. To estimate the mask that can determine the contribution of each source in the mixed signal. We adopt the magnitude spectrogram of the mixed-signal, using (STFT) with Hann window the size is set to (512)), and the reference masks of the clean sources (male-female). The network trained to predict the mask closest to the reference masks by minimizing the mean squared error between the predicted and the reference mask, the predicted mask is then used for separation. The (DRNN-BLISTM) model consists of two sub-networks, the first network is used to estimate masks for the male signal, while the other is used to estimate masks for the female signal. Each sub-network consists of three hidden layers (BLSTM) (128cell),followed by (dense layer) (64cell),and

(BLSTM) layer with (512cell). The activation function is used to train the model type (Tanh). The architecture of the proposed model is shown in figure (1).

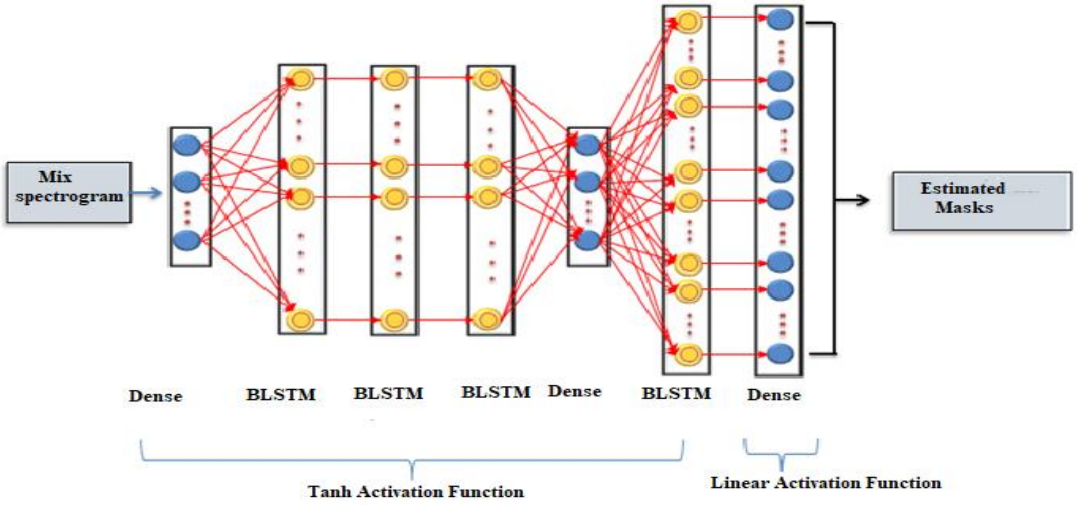


Figure (1) DRNN-BLSTM Model

6. Dataset and Results

The dataset used to train the network consists of (500) preprocessed mixed-signal consists of two different signals (female-male) with (3) seconds length and frequency (8192Hz). These audio files are collected from the internet and combined to form the mixed signals. The trained model was tested on (10) mixed signals for each mask and the results of the separation process differ according to the used mask. So the efficiency of the network performance was measured using three types of metrics: signal to distortion ratio (SDR), the signal to interference ratio (SIR) addition to signal to artifact ratio (SAR) in the signal generated by the separation process. Figure (2) shows the combined signal and the source signal, as well as the separated signal using the (IRM) mask, and Figure (2) shows the separation process for the same signal using the (ORM) mask. The performance of the separation model is evaluated based on SDR, SIR and SAR calculated using the BSS evaluation toolbox, Table(1), show the results of these three metrics when use (IRM) and (ORM) in separation task. The separating results show that the signal separation process using the (ORM) mask achieved (SDR) slightly higher (0.183.db) than it is when using the (IRM). As well as for (SIR) by (0.198 .db) is higher when using the (IRM). While (SAR) in both (IRM) and (ORM), its percentage is high, but the difference between the two masks is (0.13 .db). The efficiency of separation using (ORM) is high in the case of a high correlation between the combined signals, and the loss function represented by the learning curve of the neural network using the root mean square. Propagation (RMSprop) optimizer reduces the value of the error function to reach the best value in which the error function is lower. As in Figure (3), which shows the loss function when training the network to predict each of the filters (IRM) and (ORM) using the hyperbolic tangent function as activation function with a range of [1, -1]. We achieved the best results for both masks

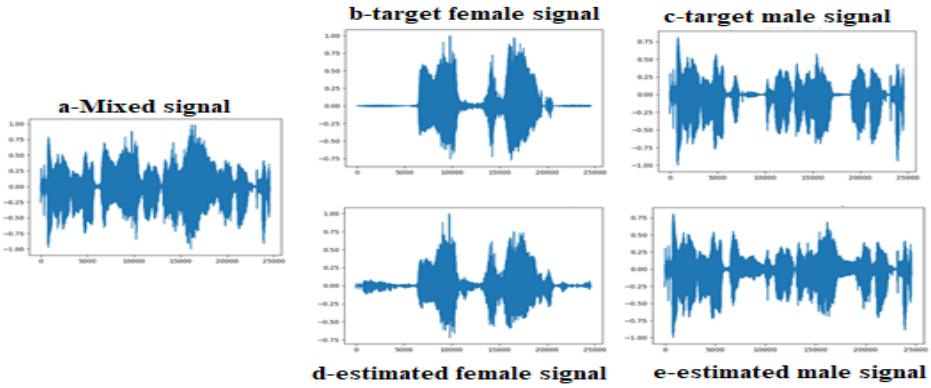


Figure (2) the original source signal and the estimated signal Using IRM Mask

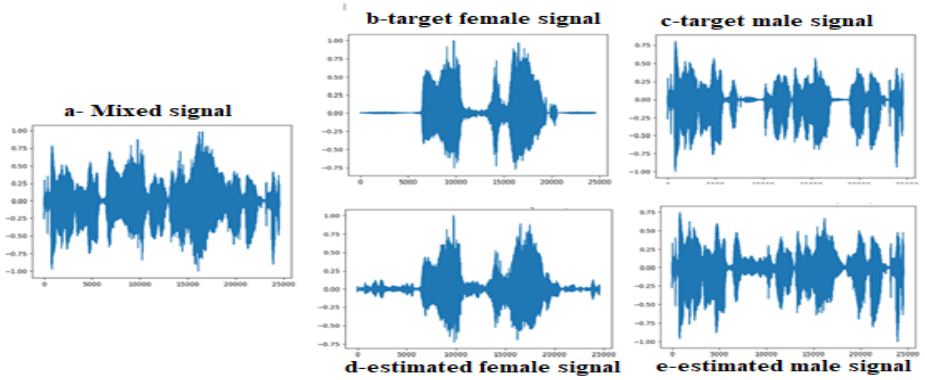


Figure (3) the original source signal and the estimated signal Using ORM Mask

Table(1) performance comparison of two masks measures

Masks	SDR	SIR	SAR
(IRM)	6.164	7.189	14.020
(ORM)	6.347	7.387	14.150

.

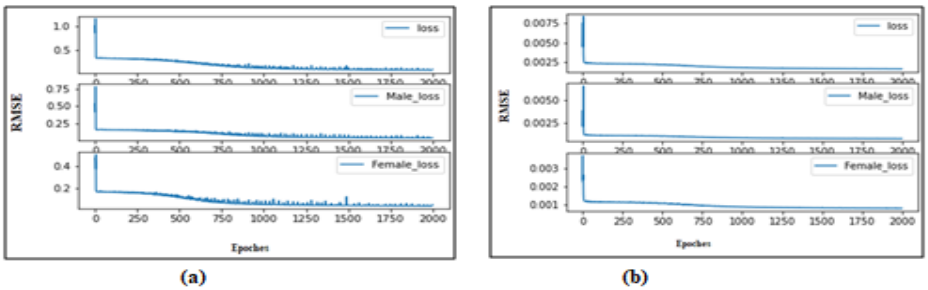


Figure (4) loss function (a) (IRM) (b) (ORM) using (Tanh)

7. Conclusions

The use of Supervised (DNN) in monaural source separation has effectively contributed to achieving separation results, especially in the case of separating large data such as audio signals. As the network extracts the characteristics from the data automatically the (TF masks) showed improvement in separation results better than the use of direct mapping, where the resulted mask was directly used to perform the separation. The (IRM) mask depend on the information of the magnitude spectrogram for both mix and source signal, the results of separation using (IRM) mask has improved good performance and the estimated signal improves high quality and intelligibility, but less than the efficiency when using (ORM) mask due to the correlation between the source signal and signals in the mixed signal.

References

[1] Rolet , Antoine ; Seguy, Vivien ; Blondel , Mathieu; awada , Hiroshi; 2018 ; **"Blind Source Separation with Optimal Transport Non-negative Matrix Factorization"**; (EURASIP) Journal on Advances in Signal Processing; No.53; https://doi.org/10.1186/s13634-018-0576-2.

[2] Bhargava , S.; 2017; **" Vocal source separation using spectrograms and spikes, applied to speech and birdsong"**, M.sc. thesis ; ETH Zurich, india .https://doi.org/10.3929/ethz-b-000175085.

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

[4] Gao, W. L. Woo, and S. S. Dlay, 2011 , "Adaptive sparsity nonnegative matrix factorization for single-channel source separation" , IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 5, pp. 989–1001.

[5] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, 2015 ,"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.

[6] E. M. Grais ; M. D. Plumbley ; 2017; **" Single Channel Audio Source Separation Using Convolutional Denoising Autoencoders"**; 5th IEEE Global Conference On Signal And Information Processing (Globalsip); Pp.1265-1269; arXiv:1703.08019V2 .

[7] Gang , Arpita ; Biyani , Pravesh; Soni, Akshay ; 2018 ; **" Towards Automated Single Channel Source Separation using Neural Networks"** ; Interspeech; pp. 3494-3498 ; arXiv:1806.08086

[8] E. M. Grais , ; Roma , Gerard ; J.R. Simpson , Andrew ; Plumbley, D. Mark; 2017;**" Single Channel Audio Source Separation using Deep Neural Network Ensembles"**; 140th Audio Engineering Society Convention; No. 9494; pp: 236–246.

[9] Yulitaa , N. Intan; Fananya , I. Mohamad ; Arymuthya , M. Aniati; 2017; **"Bi-directional Long Short-Term Memory using Quantized data of Deep Belief Networks for Sleep Stage Classification"**; 2nd International Conference on Computer Science and Computational Intelligence; (ICCSCI) ; Procedia Computer Science Vol. 116 ; pp: 530–538.

[10] Brueckner , Raymond ; Schuller, Bj¨Orn ; 2014; **" Social Signal Classification Using Deep BLSTM Recurrent Neural Networks"**; Proceedings 39th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP; pp.4856-4859; Florence; Italy .

[11] Kolbæk, Morten; Yu , Dong; Tan,Zheng-Hua; Jensen,Jesper; 2017," **Multi-Talker speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks"** , IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 10, pp. 1901-1913.

[12] Ray, Anupama ; Rajeswar, Sai; Chaudhury, Santanu; 2015; **"Text Recognition using Deep BLSTM Networks"** ; Eighth International Conference on Advance in Pattern Recognition (ICAPR) ; pp. 1- 6 ; DOI: 10.1109/ICAPR.2015.7050699.

[13] Fan , Zhe-Cheng ; Lai , Yen-Lin ; Jang ,R. Jyh-Shing ; 2017 ; **"SVSGAN: Singing Voice Separation Via Generative Adversarial Network"**; IEEE International Conference on Acoustics ; Speech and Signal Processing (ICASSP); **DOI:** 10.1109/ICASSP.2018.8462091.

[14] Xia , Shasha ; Li , Hao; Zhang, Xueliang; 2017; **"Using Optimal Ratio Mask as Training Target for Supervised Speech Separation"**; IEEE; Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC); arXiv:1709.00917.