INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Singing Voice Extraction with Attention-based Spectrograms Fusion

Hao Shi[1], *Longbiao Wang*[1*], *Sheng Li*[2*], *Chenchen Ding*[2], *Meng Ge*[1], *Nan Li*[1], *Jianwu Dang*[1,3], *Hiroshi Seki*[4]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[4]Huiyan Technology (Tianjin) Co. Ltd., Tianjin, China

{hshi_cca, longbiao_wang, gemeng, tju_linan}@tju.edu.cn,
{sheng.li, chenchen.ding}@nict.go.jp, jdang@jaist.ac.jp, hseki@huiyan-tech.com

Abstract

We propose a novel attention mechanism-based spectrograms fusion system with minimum difference masks (MDMs) estimation for singing voice extraction. Compared with previous works that use a fully connected neural network, our system takes advantage of the multi-head attention mechanism. Specifically, we 1) try a variety of embedding methods of multiple spectrograms as the input of attention mechanisms, which can provide multi-scale correlation information between adjacent frames in the spectrograms; 2) add a regular term to loss function to obtain better continuity of spectrogram; 3) use the phase of the linear fusion waveform to reconstruct the final waveform, which can reduce the impact of the inconsistent spectrogram. Experiments on the MIR-1K dataset show that our system consistently improves the quantitative evaluation by the perceptual evaluation of speech quality, signal-to-distortion ratio, signal-to-interference ratio, and signal-to-artifact ratio.

Index Terms: singing voice extraction, spectrograms fusion, attention mechanism, minimum difference masks

1. Introduction

With the progress of digital music technology and the development of streaming media, ordinary music fans are now capable of doing what used to be done only by musicians and other professionals in the music industry. Singing voice extraction has broad applications and has attracted the attention of many researchers [1]. The separated vocal contains information such as melody, lyrics, singer, and emotion, while the separated accompaniment contains information such as chord sequence, beat, and instrument [2]. Singing voice extraction can be regarded as an audio-specific source separation system [3, 4], which just extracts vocal or accompaniment from a recording of one singing voice. From this respect, it is similar to speech enhancement [5]. So singing voice extraction and speech enhancement can share a lot of approaches [4].

In recent years, supervised singing voice extraction approaches show great nonlinear mapping capability [2, 6]. Moreover, there are few or no hypotheses. These advantages attract more attention. Mapping and masking targets are two kinds of learning targets used in a supervised singing voice extraction system [4]. Mapping targets correspond to the spectral representations of clean speech [7, 8], while masking targets describe the time-frequency (T-F) relationship of clean speech to background interference [9]. Many kinds of research have been conducted through these two learning objectives. Moreover, the mapping and masking approaches have different enhancement effects in different scenarios shows some complementarity [10].

With these complementarities, a nonlinear spectrograms fusion system [11] fuses the T-F bins with the smallest distance between enhanced and clean spectrograms into one spectrogram. Although it has improved the performance of speech enhancement, there are still some problems. First, the use of minimum difference masks (MDMs) [11] to fuse the best parts of spectrograms into a new one may disrupt the data distribution of the spectrogram predicted by the neural network, resulting in discontinuity of the spectrogram. Second, multiple spectrograms can be obtained in the fusion process. Still, it does not use these spectrograms to get new phase information to replace the original noisy phase, which will most likely lead to an inconsistent spectrogram [12, 13, 14].

To overcome these problems and further improve its performance, we design an attention-based fusion system:

1) In order to obtain better continuity of spectrogram, we add a regular term to loss function.

2) In order to alleviate inconsistent spectrogram, we use the phase of the linear fusion waveform to reconstruct the final waveform, because the iterative signal reconstruction can produce better resynthesized speech [15].

3) In order to get better neural network modeling capabilities, attention mechanism is adopted. We have tried a variety of embedding [16] methods of multiple spectrograms as the input of attention mechanism [17, 18, 19].

The rest of this paper is organized as follows. Section 2 describes the minimum difference masks. Section 3 describes our proposed methods. Section 4 describes the data and experimental evaluations. A summary of the current work and outline of future work are given in Section 5.

2. Minimum difference masks

We define the distance between each separated T-F bin and its corresponding label as d_i, where spc_i denotes an enhanced spectrogram. The i in this study is $mapping$ or $masking$.

$$d_i(t,f) = |spc_i(t,f) - spc_c(t,f)| \tag{1}$$

Minimum difference masks (MDMs) [11] are to classify the T-F bins, which are nearest the labels in the multiple spectrograms. The labels of the MDMs are defined as Eq. (2). Furthermore, MDM estimation can be treated as a supervised problem

*Corresponding author.

http://dx.doi.org/10.21437/Interspeech.2020-1043

with labeled data. $\widetilde{\text{MDM}}_i(t,f)$ is set to 1 when $d_i(t,f)$ is at a minimum, and 0 otherwise. Because the spectrogram is continuous, the MDMs in the testing are real values in (0, 1). The process of the computing labels of MDMs is shown in Fig. 1.

$$\widetilde{\text{MDM}}_i(t,f) = \begin{cases} 1, & i = \arg\min_i d_i(t,f) \\ 0, & otherwise \end{cases} \tag{2}$$

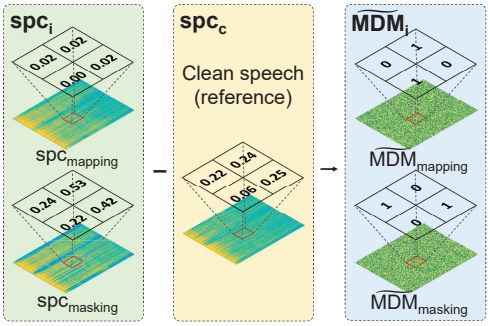


Figure 1: *Process of computing the labels of the MDMs: spc_c is the clean spectrogram, spc_i are enhanced spectrograms from the first stage, and $\widetilde{MDM}_i$ are labels of the MDMs.*

3. Proposed methods

Fig. 3 shows the procedure of spectrograms fusion. Spectrograms fusion procedure contains two stages. In the first stage, mapping and masking targets are learned in a single model with two outputs [10].

$$L_{MTL} = L_{mapping} + \alpha L_{masking} \tag{3}$$

where $L_{masking}$ and $L_{mapping}$ are computing the loss of mean squared error (MSE) gained between estimated spectrogram and the target clean spectrogram. We estimate the MDMs from the spectrograms as a supervised problem with labeled data [11].

$$\begin{aligned} L_{MDM} = & \sum_i \sum_{t,f} \left(\text{MDM}_i(t,f) - \widetilde{\text{MDM}}_i(t,f) \right)^2 \\ & + \beta \left(L_{masking} + L_{mapping} \right) \end{aligned} \tag{4}$$

where MDM_i denotes the estimated MDMs. A variety of embedding methods of multiple spectrograms as the input of attention mechanism are shown in Fig. 2. With MDMs, we could get a nonlinear fusion of spectrograms [11].

Nonlinear selection processing is conducted in the testing stage using Eq. (5), where $select_i$ denotes the nonlinearly selected portion in spc_i.

$$\text{select}_i(t,f) = \text{MDM}_i(t,f) * spc_i(t,f) \tag{5}$$

We recombine each selected portion to get the final fused spectrogram:

$$spc_f = \sum_i \text{select}_i \tag{6}$$

where spc_f denotes the final fused spectrogram.

Finally, we use the nonlinear fused spectrogram and the phase from the linear fusion constructed waveform to reconstruct the final enhanced waveform.

3.1. A regular term for MDMs-based spectrograms fusion

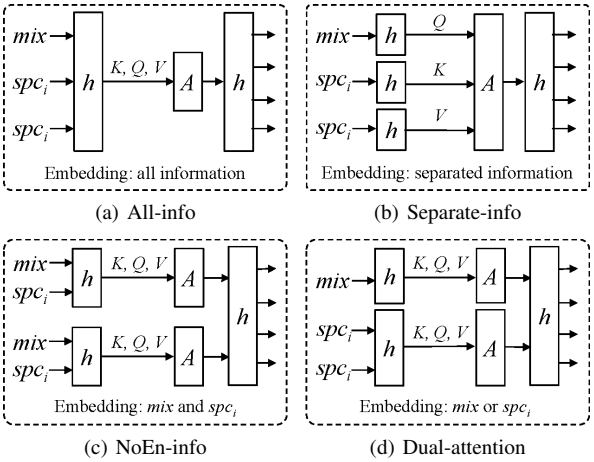
Considering the continuity of fusion spectrogram, we add an item in the process of learning:

$$L_{MDM-tend} = L_{MDM} + \gamma \left(spc_f - spc_c \right)^2 \tag{7}$$

We call the model trained using Eq. (7) as **MDM-tend**.

3.2. Embedding

Network embedding aims to map the input data into a latent space [16], so it is another representation of the input data. Besides, different input data or different network embedding methods may have a significant impact on the effectiveness of embedding. In this paper, we use a hidden layer as an embedding network.



(a) All-info (b) Separate-info

(c) NoEn-info (d) Dual-attention

Figure 2: *A variety of embedding methods of multiple spectrograms as the input of attention mechanism: mix is the noisy spectrogram; $spc_i (i \in (mapping, masking))$ is enhanced spectrograms; h is a hidden layer and A is an attention mechanism; K, Q and V are key, query and value in attention mechanism; (a) All the information as an embedding (All-info), (b) The information as an embedding separately (Separate-info), (c) The noisy and enhanced information as an embedding (NoEn-info), (d) The noisy and enhanced information modeled separately (Dual-attention).*

3.3. Attention mechanism

An attention mechanism can be described as computing a weighted sum of values, where the weight assigned to each value (V) is computed by a compatibility function of the query (Q) with the corresponding key (K). We compute the attention mechanism on a set of queries packed together into a matrix Q. The keys and values are also packed together into matrices K and V. The matrix of the outputs are computed as follows:

$$Attention\left(Q,K,V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{8}$$

where d_k is the dimension of the queries and keys.

3.4. Signal reconstruction

The iterative signal reconstruction can produce better-resynthesized speech [15], so we use the phase from the linear

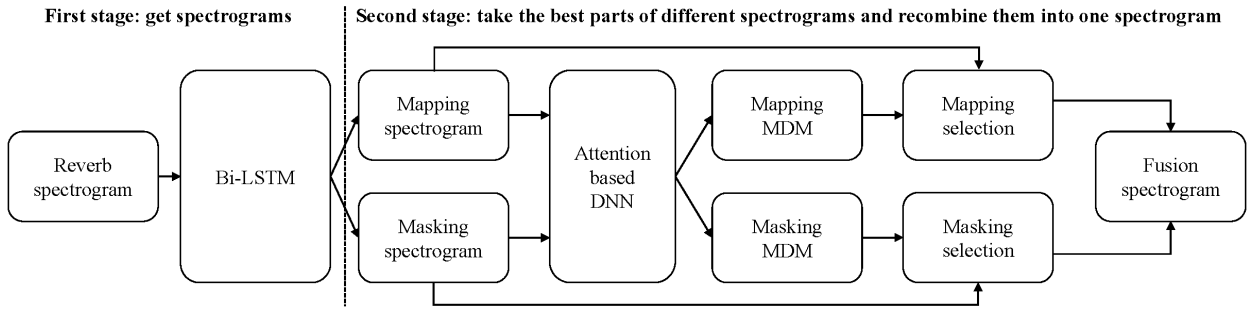


Figure 3: *Process of spectrograms fusion.*

fusion waveform and the nonlinear fused spectrogram to reconstruct the final enhanced waveform. The linear fused spectrogram (also called ensemble) obtained the following [10], where $spc_{mapping}$ and $spc_{masking}$ are two separated enhanced spectrograms.

$$spc_{LSF} = (spc_{mapping} + spc_{masking}) \,/\, 2 \tag{9}$$

4. Experiments

The experiments were conducted on the MIR-1K dataset [1] [20]. The MIR-1K dataset contains 1000 song clips recorded at a 16-kHz sampling rate with a 16-bit resolution. These clips contain mixed tracks and music accompaniment tracks, consisting of the voices of 8 females and 11 males. We selected all of the tammy's clips as the test set, a total of 8 clips. Twelve clips were selected randomly as the validation set, and the remaining 980 clips were used as the training set. We synthesized two tracks to produce monaural mix singing voice data such that the signal-to-noise ratio was equal to 0. All networks were implemented based on Tensorflow. The model's parameters were randomly initialized. The network parameters are shown in Table 1. Because the mapping and masking are both important, α was set to 1. The difference between β and γ in the interval (0, 1) has little effect on the result, so they were set as 1 and 0.5, respectively.

Table 1: *Parameters of the spectrograms fusion system.*

Settings	First stage	Second stage
Neural network	Bi-LSTM	Attention + DNN
Hidden layers	2	1
Nodes per layer	512	1024
Input dimension	257	257 * 3
Output dimension	257 * 2	257 * 4
Learning rate	0.01	0.01
Epoch	30	30
Batch size	8	8

The perceptual evaluation of speech quality (PESQ) [21], signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) were used as evaluation metrics [22]. "S-Masking" denotes the masking approach of using $L_{masking}$ as training loss, while "S-Mapping" denotes the mapping approach of using $L_{mapping}$ as training loss. "M-Mapping" and "M-Masking" denote two outputs of the MTL approaches using Eq. (3). "M-LSF" denotes the approach using Eq. (11) [10]. "uPIT-vocal" denotes the vocal output which trained using uPIT [23]. As can be seen from Table 2. Mapping and masking approaches have different effects on measures; e.g., S-Mapping yielded better results than S-Masking for the PESQ, SDR, and SIR, while the results for the SAR were the opposite. The multi-targets learning approaches outperformed single learning approaches; i.e., the M-Mapping and M-Masking showed consistently superior measures. One of the multi-targets learning model outputs was consistently superior to the other; i.e., the M-Masking performed better than the M-Mapping. The uPIT-vocal approach showed a strong ability to singing voice separation, but there was a drop in PESQ.

[1] https://sites.google.com/site/unvoicedsoundseparation/mir-1k

4.1. The effect of regular terms and phase

Several observations could be made in Table 2. "MDM-tend" denotes nonlinear fusion method using Eq. (15) and "+phase" denotes phase of linear fusion waveform were used when reconstructing waveform. Adding regular terms, change information of the spectrogram, to the neural network gives better results; e.g., the MDM-tend approach outperformed the MDM approach. A better phase can be obtained by extracting the phase in the speech of the linear fusion approach; this yielded an average PESQ gain of 0.052, an average SDR gain of 0.328, an average SAR gain of 0.206, and average SIR gain of 0.554.

Table 2: *Results of nonlinear spectrogram fusion approaches.*

Systems	SDR	SAR	SIR	PESQ
Mix signal	0.058	140.81	0.058	1.112
S-Masking	9.315	11.645	13.448	1.629
S-Mapping	9.324	11.496	13.743	1.914
M-Mapping	9.215	11.261	13.835	1.965
M-Masking	9.804	11.834	14.425	1.851
M-LSF [10]	9.770	11.934	14.161	2.090
uPIT-vocal [23]	9.751	11.902	14.141	1.854
MDM [11]	10.036	11.830	15.096	2.217
MDM-tend	10.063	11.848	15.142	2.212
+phase	**10.391**	**12.050**	**15.709**	**2.263**

4.2. The effect of attention mechanism

Several observations could be made in Table 3. The attention mechanism helped to model the relationships between the spectrograms, which in turn reduced the degree of distortion and interference of the speech. Moreover, through the fluctuation

of evaluation metrics, it can be seen that the attention mechanism modeling can better reduce additive noise and musical noise. However, the effect of accompaniment was less reduced. MDM-tend-Separate-info showed the best performance; this means that the attention mechanism can better learn the information from the embedding of a single spectrogram. All modeling methods contributed to speech enhancement, which verified the robustness of proposed approaches. No one system can get consistent improvements in all metrics, this may mean that attention mechanisms get different information in different ways of modeling.

Table 3: *Results of different embedding methods of multiple spectrograms (+phase).*

Systems	SDR	SAR	SIR
MDM-tend	10.391	12.050	**15.709**
MDM-tend-All-info	10.397	12.100	15.626
MDM-tend-Separate-info	**10.461**	**12.252**	15.491
MDM-tend-NoEn-info	10.417	12.132	15.613
MDM-tend-Dual-attention	10.397	12.152	15.503

4.3. The attention mechanism

In this experiment, the attention had three heads; each head was a representation subspace [17]. Fig. 4. shows an attention weights example. Several observations could be made from this figure. Overall, the attention mechanism of greater weight proceeds monotonically. To some extent, this shows that the attention mechanism used in this paper worked [24]. In detail, each frame and its adjacent frames tend to have a greater weight. We use the attention mechanism similar to that in [18].

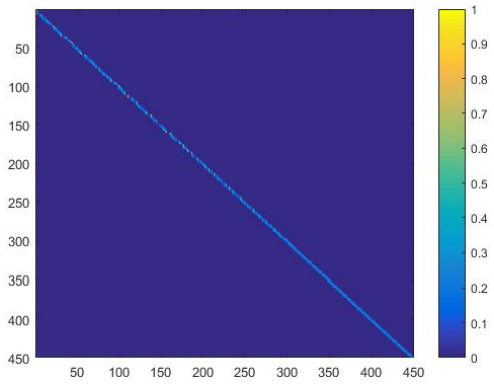
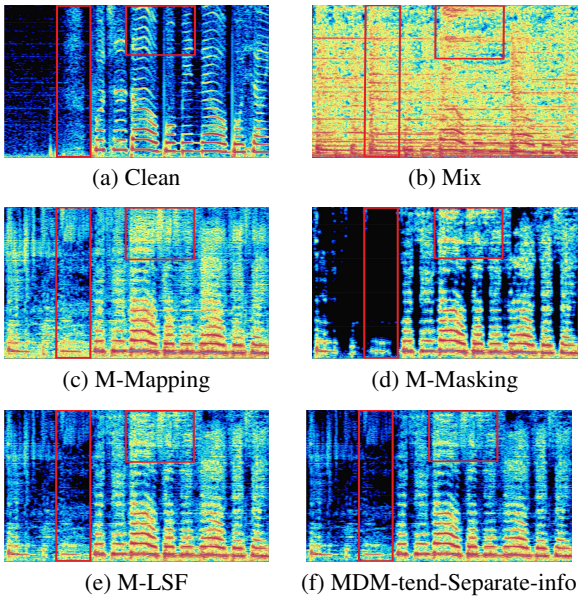


Figure 4: *An attention weights example: The vertical axis indices and the horizontal axis indices correspond to frames in the spectrogram.*

4.4. Magnitude spectrogram

Fig. 5 shows the magnitude of spectrograms. All of the enhanced approaches get most of the speech signal in the mixed-signal. However, their extracted speech signal still contains part of the accompaniment signal. All of the enhanced approaches could restore the spectrum at low frequencies buried in the mixture signal. However, they were poor at recovering the high frequencies. There was still a lot of noise in the high-frequency component of the M-Mapping spectrogram, while the high-frequency part of the M-Masking spectrogram removed too much vocal information. M-LSF made a compromise by averaging the M-Mapping and M-Masking, but the high frequencies recovering were still not good enough. Although the MDM-tend-Separate-info (+phase) approach still had some noise at high frequencies, it had some improvement over the other methods, some high-frequency details were restored in particular. This may be because in the process of fusion, high-frequency part in preference to select the masking spectrogram, part combines some information of the mapping spectrogram.



(a) Clean (b) Mix

(c) M-Mapping (d) M-Masking

(e) M-LSF (f) MDM-tend-Separate-info

Figure 5: *Magnitude of spectrograms: (a) clean, (b) mixed, (c) mapping-based, (d) masking-based, (e) linear fusion-based, and (f) MDM-tend-Separate-info (+phase).*

5. Conclusion

The minimum difference masks (MDMs) [11] had shown strong enhancement abilities, especially for SIR and PESQ. Experiments on the MIR-1K dataset show that our system consistently and significantly improves the quantitative evaluation. First, the regular term could help the system get better performance on SDR, SAR, and SIR. Second, we use the phase from the linear fusion constructed waveform to reconstruct the final enhanced waveform that can improve all the quantitative evaluation performance. Besides, different ways of embedding provide different enhancement effects, and we observed that the MDM-tend-Separate-info had the best modeling capability. The attention mechanism provided us with a new idea that finding keyframes in the spectrogram may help speech enhancement, and this is our work for the future.

6. Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grant 61771333, the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330. Sheng Li is partially supported by JSPS KAKENHI Grant No. 19K24376 and NICT tenure-track startup fund “Research of advanced automatic speech recognition technologies”, Japan.

7. References

[1] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE TASLP*, vol. 15, no. 4, pp. 1475–1487, 2007.

[2] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. ISMIR*, 2017, pp. 745–751.

[3] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE TASLP*, vol. 20, no. 4, pp. 1118–1133, 2012.

[4] Z. Rafii, A. Liutkus, F. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM TASLP*, vol. 26, no. 8, pp. 1307–1335, 2018.

[5] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[6] S. Yang and W. Zhang, "Singing voice separation based on deep regression neural network," in *Proc. ISSPIT*, 2019, pp. 1–5.

[7] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 7–19, 2015.

[8] Y. Ueda, L. Wang, A. Kai, and B. Ren, "Environment-dependent denoising autoencoder for distant-talking speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–11, 2015.

[9] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM TASLP*, vol. 22, no. 12, pp. 1849–1858, 2014.

[10] L. Sun, J. Du, L. Dai, and C. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *Proc. HSCMA*, 2017, pp. 136–140.

[11] H. Shi, L. Wang, M. Ge, S. Li, and J. Dang, "Spectrograms fusion with minimum difference masks estimation for monaural speech dereverberation," in *Proc. ICASSP*, 2020, pp. 7544–7548.

[12] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[13] Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, and M. Iwahashi, "Dnn-based amplitude and phase feature enhancement for noise robust speaker identification." in *Proc. Interspeech*, 2016, pp. 2204–2208.

[14] Z. Oo, L. Wang, K. Phapatanaburi, M. Iwahashi, S. Nakagawa, and J. Dang, "Phase and reverberation aware dnn for distant-talking speech enhancement," *Multimedia Tools and Applications*, vol. 77, no. 14, pp. 18 865–18 880, 2018.

[15] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM TASLP*, vol. 23, no. 6, pp. 982–992, 2015.

[16] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proc.SIGKDD*, 2016, pp. 1225–1234.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[18] X. Hao, C. Shan, Y. Xu, S. Sun, and L. Xie, "An attention-based neural network approach for single channel speech enhancement," in *Proc. ICASSP*, 2019, pp. 6895–6899.

[19] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, 2015, pp. 577–585.

[20] C. Hsu and J. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE TASLP*, vol. 18, no. 2, pp. 310–319, 2010.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752 vol.2.

[22] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[23] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM TASLP*, vol. 25, no. 10, pp. 1901–1913, 2017.

[24] C. Raffel, M. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *Proc. ICML*, 2017, pp. 2837–2846.