A Skip Attention Mechanism for Monaural Singing Voice Separation

Weitao Yuan, Shengbei Wang, Xiangrui Li, Masashi Unoki, and Wenwu Wang, *Senior Member, IEEE*

***Abstract*—This work proposes a simple but effective attention mechanism, namely Skip Attention (SA), for monaural singing voice separation (MSVS). First, the SA, embedded in the convolutional encoder-decoder network (CEDN), realizes an attention-driven and dependency modeling for the repetitive structures of the music source. Second, the SA, replacing the popular skip connection in the CEDN, effectively controls the flow of the low-level (vocal and musical) features to the output and improves the feature sensitivity and accuracy for MSVS. Finally, we implement the proposed SA on the Stacked Hourglass Network (SHN), namely Skip Attention SHN (SA-SHN). Quantitative and qualitative evaluation results have shown that the proposed SA-SHN achieves significant performance improvement on the MIR-1K dataset (compared to the state-of-the-art SHN) and competitive MSVS performance on the DSD100 dataset (compared to the state-of-the-art DenseNet), even without using any data augmentation methods.**

***Index Terms*—Skip attention, stacked hourglass network, monaural singing voice separation.**

I. Introduction

Music source separation (MSS) is one of the fundamental research areas for music signal processing. Monaural singing voice separation (MSVS) is an important examplar of MSS, which aims to separate the singing voice (vocals) and the background musical accompaniment from a single channel mixture signal. The traditional largely unsupervised methods provide effective frameworks for MSVS such as [1]. Recently, the data-driven method, especially the Deep Neural Network (DNN) based methods [2], [3], have emerged and provided state-of-the-art performance for MSVS. There are generally three basic structures to construct DNNs: Feed-Forward Network (FFN) [4], Recurrent Neural Network (RNN) [5], and Convolutional Neural Network (CNN) [6], [7]. Recently the RNN and CNN have been combined to improve the MSS [8], [9].

Since the CNN is effective for feature extraction in time-frequency (T-F) domain, the state-of-the-art methods usually employ the convolutional encoder-decoder networks (CEDNs) for MSVS, e.g., the U-net [10] and the Stacked Hourglass Network (SHN) [11]. In these CEDNs, the input spectrogram is compressed (by the encoder) into a bottleneck layer to obtain a lower dimensional descriptor and then the descriptor is re-expanded to the size of the target spectrogram (by the decoder) [10]. In addition, with the help of additional skip connections (or similar structures), the CEDNs can recreate fine and low-level details for high-quality MSVS. However, in spite of their popularities, several issues need to be addressed to improve the current CEDNs.

First, music relies heavily on its repetitions to build the logical structure and meaning [12]. These repetitions made of recurring elements may appear at various levels of the music, e.g., from very basic elements (individual notes, timber, or pitch) to larger structures (e.g., chords) [13]. These multi-scale repetitions effectively distinguish the musical accompaniment from the vocals which are less redundant and mostly harmonic [14]. Therefore, effectively modeling the repetitive structures in the mixture signal would be a promising solution for DNN based MSVS. Since the repetitive structures can be observed as the similarities between different regions in the T-F representations (e.g., magnitude spectrogram), the MSVS methods need to attend to the different T-F regions in order to capture the dependencies across different frequencies in the mixture. However, the convolution operator that has a local receptive field can only model the repeating patterns locally. To deal with this problem, current CEDNs try to pass the input (mixture) through multiple convolutional layers, to form a cascade framework for MSVS. Unfortunately, according to [15], the optimization algorithms used in these CEDNs are usually not effective in capturing the dependencies across multiple layers, especially for those complicated repetitive musical structures.

Second, it is well known that even a minor linear shift in the T-F representations could introduce significant distortions on vocal and music perception [10]. Thus current CEDNs usually have skip or similar connections to pass the low-level vocal and musical features from the input to the output, to obtain high-level precise details for the estimated sources [10], [11]. Such direct connections, however, also allow the input features to circumvent the screening of the bottleneck layer, which is a necessary step to extract the essential singing voice and music features through dimensionality reduction. As a result, these direct (or skip) connections may weaken the encoder-decoder bottleneck structure and degrade the separation performance.

To address the above issues, this work proposes a Skip Attention (SA) mechanism for MSVS by including a post-processing layer in the CEDN to refine the multiresolution encoder/decoder

Manuscript received June 20, 2019; revised August 11, 2019; accepted August 13, 2019. Date of publication August 19, 2019; date of current version September 3, 2019. This work was supported by Natural Science Foundation of Tianjin (17JCQNJC00100 and 19JCYBJC15600) and National Natural Science Foundation of China (61902280). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Odette Scharenborg. *(Corresponding author: Shengbei Wang.)*

W. Yuan, S. Wang, and X. Li are with the School of Computer Science and Technology, Tianjin Polytechnic University, Tianjin 300160, China (e-mail: weitaoyuan@hotmail.com; wangshengbei@tjpu.edu.cn; 1611650430@stu.tjpu.edu.cn).

M. Unoki is with the School of Information Science, Japan Advanced Institute of Science and Technology, Nomi 923-1292, Japan (e-mail: unoki@jaist.ac.jp).

W. Wang is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: w.wang@surrey.ac.uk).

Digital Object Identifier 10.1109/LSP.2019.2935867

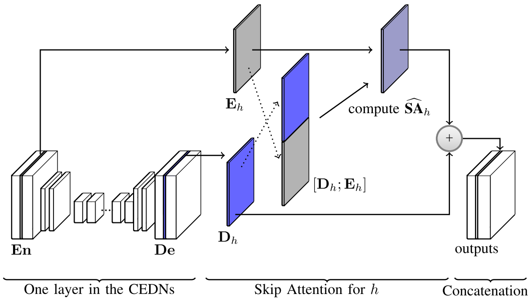

Fig. 1. The Skip Attention (SA) mechanism for one layer in the CEDN.

structure and control the information flow between the encoding and decoding layers. First, the SA, as a non-local operation in the multiresolution CEDN, can use cues from all time-frequency regions to select the relevant features from both the input and output of the CEDN and hence efficiently model the multi-scale repetitive structures of the music source. Second, the proposed SA also introduces an inter-attention mechanism to replace the direct or skip connections between the input and the output in current CEDNs. This mechanism allows the input and reconstructed output to interact with each other and thus effectively controls the input low-level (vocal and musical) information flow to the output: each feature will be enhanced or suppressed through this inter-attention. Thus, this inter-attention mechanism will selectively allow the relevant low-level details of the vocal and music sources to pass through, and meanwhile, keep the bottleneck structure of CEDNs working effectively to extract essential vocal or musical features.

One of the prior works related to this work is the Transformer [16], which includes the "encoder-decoder attention" layers: the queries come from the decoder, and the keys and values come from the encoder. Similar designs can be found in [17]–[19] for machine translation. However, compared to these works, the proposed SA utilizes a different "encoder-decoder attention" mechanism: we use queries from the encoder, and the keys and values are constructed by concatenating both the outputs of the decoder and the inputs of the encoder. This special mechanism is designed to achieve two goals: (i) modeling the repetitive structures in the musical source; (ii) controlling the flow of the input low-level (vocal and musical) information to the output by attending to both the decoder and encoder. Besides, instead of replacing the convolutional structure with attention mechanisms as the transformer does, we improve the bottleneck structure of the CEDNs by replacing its skip connection with the SA (that is the reason why we name the proposed mechanism as 'skip' attention).

II. The Proposed Method

A. Skip Attention Mechanism

In an encoder-decoder structure shown in Fig. 1, we assume that the input is an internal time-frequency feature $\mathbf{En} \in \mathbb{R}^{N \times T \times d}$, and the reconstructed output is $\mathbf{De} \in \mathbb{R}^{N \times T \times d}$, where N is the number of frequency bins, T is the number of time frames, and d is the number of channels. To capture the dependency across frequencies, we fix the h to compute the attention matrix in frequency-channel. Specifically, for each h $(1 \leq h \leq T)$, the two inputs for SA are

$$\mathbf{E}_h = \mathbf{En}[:, h, :] \in \mathbb{R}^{N \times d}, \mathbf{D}_h = \mathbf{De}[:, h, :] \in \mathbb{R}^{N \times d}. \quad (1)$$

The specific steps for computing SA are:

1) Linear Projection: At first, the input query, key, and value matrices for computing the skip attention are defined as $\mathbf{E}_h$, $[\mathbf{D}_h; \mathbf{E}_h]$, and $[\mathbf{D}_h; \mathbf{E}_h]$, respectively, where the semicolon represents the vertical concatenation of matrices. Then we apply the projection matrix $\mathbf{W}^Q$, $\mathbf{W}^K$ and $\mathbf{W}^V \in \mathbb{R}^{d \times d}$ to the query, key and value respectively to compute their projected versions, i.e.,

$$\mathbf{E}_h^Q = \mathbf{E}_h \mathbf{W}^Q \in \mathbb{R}^{N \times d}, \quad (2)$$

$$[\mathbf{D}_h^K; \mathbf{E}_h^K] = [\mathbf{D}_h; \mathbf{E}_h] \mathbf{W}^K \in \mathbb{R}^{2N \times d}, \quad (3)$$

$$[\mathbf{D}_h^V; \mathbf{E}_h^V] = [\mathbf{D}_h; \mathbf{E}_h] \mathbf{W}^V \in \mathbb{R}^{2N \times d}. \quad (4)$$

2) Skip Attention: We use the projected query $\mathbf{E}_h^Q$ to produce a scaled dot-product attention matrix (denoted as $\mathbf{V}_{\text{att},h}$) over the projected keys $[\mathbf{D}_h^K; \mathbf{E}_h^K]$,

$$\mathbf{V}_{\text{att},h} = \frac{\mathbf{E}_h^Q [\mathbf{D}_h^K; \mathbf{E}_h^K]^T}{\sqrt{d}} \in \mathbb{R}^{N \times 2N}. \quad (5)$$

The attention matrix $\mathbf{V}_{\text{att},h}$ will be processed with masking and softmax-function [16] (denoted as Softm in Eq. (6)) to form the skip attention $\mathbf{SA}_h$ for h,

$$\mathbf{SA}_h = \text{Softm}\left(\mathbf{V}_{\text{att},h}\right) [\mathbf{D}_h^V; \mathbf{E}_h^V] \in \mathbb{R}^{N \times d}. \quad (6)$$

Intuitively, the nonlinear softm in $\mathbf{SA}_h$ operates on matrix $\mathbf{V}_{\text{att},h}$, which includes two types of attention: (1) the inter-attention between the input $\mathbf{E}_h$ and output $\mathbf{D}_h$; (2) the self-attention of the input $\mathbf{E}_h$, that is,

$$\mathbf{V}_{\text{att},h} = \frac{\mathbf{E}_h^Q [\mathbf{D}_h^{K^T}, \mathbf{E}_h^{K^T}]}{\sqrt{d}} = \frac{[\overbrace{\mathbf{E}_h^Q \mathbf{D}_h^{K^T}}^{(1)}, \overbrace{\mathbf{E}_h^Q \mathbf{E}_h^{K^T}}^{(2)}]}{\sqrt{d}}. \quad (7)$$

According to Eq. (7), the operation softm$\left(\mathbf{V}_{\text{att},h}\right)$ in Eq. (6) considers the above two attentions as a whole and hence enables $\mathbf{SA}_h$ to select suitable features from the feature combined matrix $[\mathbf{D}_h^K; \mathbf{E}_h^K] \in \mathbb{R}^{2N \times d}$ (see Eq. (6)), where both the input and the output features of the CEDN are included. As a result, the SA layer, as a post-processing step in the CEDN, provides a flexible mechanism for selecting features from the input, the output, or both. While in the original CEDN, the input will be added directly to the output via a skip connection, i.e., all the input and output features are mixed together without effective feature selection.

Finally, we apply the residue connection and layer normalization [23] (denoted as LN in Eq. (8)) as follows,

$$\widehat{\mathbf{SA}}_h = \text{LN}\big(\mathbf{SA}_h + \mathbf{E}_h\big). \quad (8)$$

3) Addition: The normalized skip-attention $\widehat{\mathbf{SA}}_h$ and the output of the decoder $\mathbf{D}_h$ are added as the final output for h.

B. Skip Attention Stacked Hourglass Network

As an example, we use the proposed SA to enhance the popular SHN [11], which is made of several stacked Hourglass Networks (HNs). Specifically, within each HN, multiple SAs are added as the post-processing layers for multiresolution feature extraction (see Fig. 2). For a fair comparison, the architecture of the SA-SHN shown in Fig. 3 follows the SHN, where the input magnitude spectrogram is passed through the initial convolutional layers and then fed to the four SA driven Hourglass

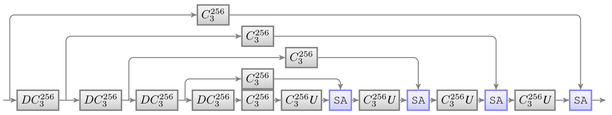

Fig. 2. The SA driven Hourglass Network (denoted as SA-HN in Fig. 3), where C_m^n is the convolutional layers with filter size of $m \times m$ and output channels of n, the D is 2×2 max pooling, and U is 2×2 upsampling.

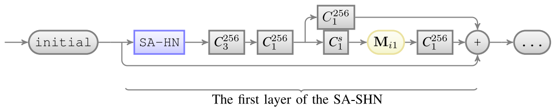

Fig. 3. The SA-SHN: the initial step includes $C_7^{64}C_3^{128}C_3^{128}C_3^{128}C_3^{256}$, where C_m^n is the convolutional layers with filter size of $m \times m$ and output channels of n; the SA-HN represents the SA driven Hourglass Network in Fig. 2; s represents the number of separated sources; $\mathbf{M}_{ij}$ $(1 \leq i \leq s)$ represents the i-th source mask for the j-th layer; ... presents the second or more layer.

Networks to output masks for each source. In Fig. 3, each layer (with the output mask $\mathbf{M}_{ij}$ for the i-th source in the j-th layer, $1 \leq i \leq s$) can be iterated (stacked) many times and s is the number of sources to separate (e.g., $s = 2$ for MSVS). The loss function $L_{1,1}$ norm in [11], [24] is adopted for a fair comparison. Formally, given the magnitude $\mathbf{X}$ of the mixture, the i-th ground truth source $\mathbf{Y}_i$, and the predicted mask for the i-th source in the j-th module $\mathbf{M}_{ij}$, the loss function [11] is defined as, $\mathcal{J} = \sum_{i=1}^{s} \sum_{j=1}^{D} \|\mathbf{Y}_i - \mathbf{X} \odot \mathbf{M}_{ij}\|_{1,1}$, where $\odot$ denotes element-wise multiplication of the matrix and D is the number of the SA driven Hourglass Networks.

III. Experiments

We followed the experimental settings in [11] to evaluate the proposed SA-SHN for MSVS and MSS. The input was the magnitude spectrogram of the mixture calculated by Short-Time Fourier Transform (STFT) with a window size of 1024 and a hop size of 256. The SA-SHN was trained to predict soft masks for s sources. By multiplying the predicted masks with the magnitude spectrogram of the mixture, the estimated magnitude spectrogram of each source was obtained. The time domain sources were obtained via inverse STFT applied to each estimated magnitude spectrogram and the phase spectrogram of the mixture.

The performance of different methods was evaluated on two popular datasets: MIR-1K [25] and DSD100 [26]. For a fair comparison on the MIR-1K, we followed the evaluation conditions in [11], [20], [21]: 175 clips performed by 'abjones' and 'amy' were used for training and the other 825 clips were used for testing. For MSVS on the DSD100, we followed [11] to convert all sources to monophonic and then we added three sources except for the vocals together to form the musical components (Acc.) source.

Both the proposed SA-SHN and the baseline SHN were trained using Adam optimizer [27] with an initial learning rate of 10^{-4} and a batch size of 1. We trained these two networks with 60,000 iterations for the MIR-1K dataset and with 600,000 iterations for the DSD100 dataset, and the learning rate is decreased to 2×10^{-5} when 80% of the training is finished. No data augmentation was applied during training for both datasets as in [11]. The computational efficiency of the SA-SHN is slightly lower than the SHN, e.g., in four-layer case, the training/testing took approximately 3 hours/15 minutes for the SHN and 5 hours/20 minutes for the SA-SHN on the MIR-1K dataset using a single GPU (GeForce GTX 1080 Ti). For quantitative evaluation, the separation performance was measured by BSS-EVAL toolkit [28] with respect to three criteria, i.e., source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR). We also use Normalized SDR (NSDR) [29], Global SIR (GSIR), Global SAR (GSAR), and Global NSDR (GNSDR) [11], [21] to evaluate the results.

TABLE I
Quantitative Results for MSVS on MIR-1K (in dB)

Singing voice (Vocal)			
Method	GNSDR	GSIR	GSAR
MLRR [20]	3.85	5.63	10.70
DRNN [21]	7.45	13.08	9.68
ModGD [22]	7.50	13.73	9.45
U-Net [10]	7.43	11.79	10.42
SHN-4 [11]	10.60	15.92	12.69
SA-SHN-1	11.07	16.62	13.00
SA-SHN-2	11.31	17.16	13.11
SA-SHN-4	**11.66**	**17.65**	**13.38**
Musical accompaniment (Acc.)			
Method	GNSDR	GSIR	GSAR
MLRR [20]	4.19	7.80	8.22
U-Net [10]	7.45	11.43	10.41
SHN-4 [11]	9.99	14.52	12.47
SA-SHN-1	10.04	14.03	12.91
SA-SHN-2	10.28	14.42	13.00
SA-SHN-4	**10.60**	**14.92**	**13.17**

TABLE II
Statistics of MSVS Performance on MIR-1K (in dB), Where SA-4 Represents the SA-SHN-4

Vocal	NSDR		SDR		SIR		SAR	
	SHN-4	SA-4	SHN-4	SA-4	SHN-4	SA-4	SHN-4	SA-4
Med.	10.71	**11.89**	10.79	**12.00**	16.08	**17.88**	12.58	**13.51**
MAD	1.99	**1.87**	2.04	**1.88**	2.94	**2.63**	**1.57**	1.63
Mean	10.58	**11.66**	10.67	**11.74**	15.87	**17.61**	12.71	**13.40**
SD	3.09	**3.00**	3.10	**3.02**	4.53	**4.30**	**2.36**	2.50
Acc.	NSDR		SDR		SIR		SAR	
	SHN-4	SA-4	SHN-4	SA-4	SHN-4	SA-4	SHN-4	SA-4
Med.	10.15	**10.83**	10.22	**10.88**	14.82	**15.26**	12.44	**13.22**
MAD	1.85	**1.77**	1.87	**1.77**	2.20	**2.17**	1.59	**1.44**
Mean	9.98	**10.59**	10.04	**10.66**	14.50	**14.89**	12.48	**13.19**
SD	2.97	**2.96**	2.95	**2.95**	**3.68**	3.70	2.53	**2.45**

A. MIR-1K Dataset

The quantitative evaluation results on MIR-1K dataset are shown in Table I. We trained the SA-SHN ($s = 2$ for MSVS) with varying layer numbers from 1 to 4. In the following we use SA-SHN-n to represent n layer SA-SHN and SHN-n to represent n layer SHN. It can be seen in Table I that our SA-SHN-1 (even with only one layer) significantly outperformed the original SHN-4 (the best previous method on the MIR-1K dataset) and the other methods for all evaluation criteria, except for GSIR for Acc. The proposed SA-SHN reached its highest performance with 4 layers.

In Table II, we compare the separation performance for all the song clips[1] obtained by the proposed SA-SHN-4 and the SHN-4 in [11]. As suggested by [30], the mean (Mean) SDR with its standard deviation (SD) alone were not sufficient to measure the vocal performance, we thus adopted the median (Med.) with its median absolute deviation (MAD), which were more robust against outliers. From Table II, it can be seen that the SA-SHN-4 was superior to SHN-4 in most statistics.

[1] We took each song clip as one unit without considering its length to compute NSDR/SDR/SIR/SAR for both the SA-SHN and SHN.

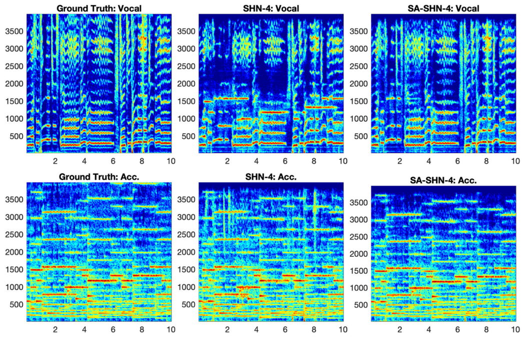

Fig. 4. Qualitative comparison of the proposed SA-SHN-4 and SHN-4 on the song clip of *khair_6_06* in MIR-1K dataset.

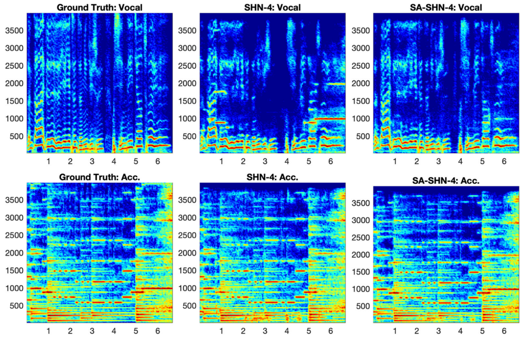

Fig. 5. Qualitative comparison of the proposed SA-SHN-4 and SHN-4 on the song clip of *stool_1_04* in MIR-1K dataset.

Qualitative results of our SA-SHN-4 and the SHN-4 are shown in Figs. 4–5. For the song clip *khair_6_06* in Fig. 4, the SA-SHN-4 recovered much more original vocal harmonics compared to the SHN-4. In particular, the vocal spectrogram of the SHN-4 around 1~2 s and 4~6 s had lost a large amount of harmonic details, while the proposed SA-SHN-4 recovered most of these fine details. This phenomenon is attributed to the SA, which successfully controls the essential low-level vocal information to flow to the output. On the other hand, for Acc. spectrogram in Fig. 5 (the song clip of *stool_1_04*), we can see that the Acc. obtained by SHN-4 around 5~6.5 s missed one important frequency component around 1000 Hz, while the SA-SHN-4 almost recaptured all the frequency components of the ground truth Acc. around 5~6.5 s, including the frequency component around 1000 Hz, which verified the effectiveness of the proposed SA.

B. DSD100 Dataset

For the MSVS task on DSD100, we set $s = 2$ to evaluate the proposed SA-SHN-1 to SA-SHN-4. In order to compare with the previous reported MSVS methods, we compared SDRs for all songs in the DSD100 dataset and then computed median values, which are listed in Table III. Compared with the state-of-the-art MM-DenseNet [35], the proposed SA-SHN-4 achieved 0.44 dB improvement for Vocals and 0.50 dB for Acc.

For the MSS task on DSD100, we set $s = 4$ to evaluate the proposed method (SA-SHN-4). It can be seen from Table IV that the SA-SHN-4 was superior to SHN-4 in most metrics. We also compare our SA-SHN-4 with some other methods, as shown in Table V. The SA-SHN-4 greatly improved the separation performance of SHN-4 for Vocals (1.04 dB gain), and slightly improvement for other sources (0.09 dB for Drums, 0.03 dB for Other and 0.01 dB for Bass). Besides, the vocal separation performance of the SA-SHN-4 is 0.21 dB higher than that of the state-of-the-art MM-DenseNet [35]. This seems to suggest that the proposed method is more effective for repetitive music and harmonic vocals, but less effective for non-repetitive sounds, such as bass and drums.

TABLE III
MEDIAN SDR VALUES (IN DB) FOR MSVS ON DSD100

Method	Vocals	Accompaniment (Acc.)
DeepNMF [31]	2.75	8.90
wRPCA [32]	3.92	9.45
NUG [33]	4.55	10.29
BLEND [34]	5.23	11.70
MM-DenseNet [35]	6.00	12.10
SHN-4 ($s = 2$) [11]	5.64	12.15
SA-SHN-1 ($s = 2$)	5.72	12.26
SA-SHN-2 ($s = 2$)	5.89	12.20
SA-SHN-3 ($s = 2$)	6.04	12.49
SA-SHN-4 ($s = 2$)	**6.44**	**12.60**

TABLE IV
MUSIC SOURCE SEPARATION PERFORMANCE ON DSD100 (IN DB)

Vocal	SDR		SIR		SAR	
	SHN-4	SA-SHN-4	SHN-4	SA-SHN-4	SHN-4	SA-SHN-4
Med.	5.17	**6.21**	12.04	**13.91**	6.73	**7.32**
MAD	**1.69**	1.76	**2.11**	2.46	**1.68**	1.69
Mean	4.90	**5.66**	11.93	**13.35**	6.51	**7.03**
SD	**3.01**	3.10	**4.28**	4.51	**2.18**	2.29
Bass	SDR		SIR		SAR	
	SHN-4	SA-SHN-4	SHN-4	SA-SHN-4	SHN-4	SA-SHN-4
Med.	1.88	**1.89**	**5.14**	5.29	6.50	**6.83**
MAD	3.24	**3.01**	**3.58**	3.84	2.14	**1.81**
Mean	2.46	**2.57**	6.07	**6.12**	6.67	**6.84**
SD	4.45	**4.21**	5.31	**5.19**	2.99	**2.65**
Drum	SDR		SIR		SAR	
	SHN-4	SA-SHN-4	SHN-4	SA-SHN-4	SHN-4	SA-SHN-4
Med.	4.24	**4.33**	10.33	**10.74**	6.11	**6.21**
MAD	**2.11**	2.20	**2.16**	2.27	2.46	**2.21**
Mean	4.10	**4.23**	10.15	**10.29**	6.19	**6.30**
SD	**3.53**	3.73	**4.13**	4.33	**3.19**	3.29
Other	SDR		SIR		SAR	
	SHN-4	SA-SHN-4	SHN-4	SA-SHN-4	SHN-4	SA-SHN-4
Med.	2.56	**2.59**	**6.64**	6.62	5.71	**6.07**
MAD	**1.15**	1.47	**1.81**	2.26	**1.01**	1.17
Mean	1.70	**2.00**	**6.07**	5.95	5.20	**5.81**
SD	3.38	**3.27**	3.82	**3.75**	2.32	**2.27**

TABLE V
MEDIAN SDR FOR MSS ON DSD100 (IN DB)

Method	Bass	Drums	Other	Vocals
dNMF [36]	0.91	1.87	2.43	2.56
DeepNMF [31]	1.88	2.11	2.64	2.75
BLEND [34]	2.76	3.93	3.37	5.13
MM-DenseNet [35]	**3.91**	**5.37**	**3.81**	6.00
SHN-4 ($s = 4$) [11]	1.88	4.24	2.56	5.17
SA-SHN-4 ($s = 4$)	1.89	4.33	2.59	**6.21**

IV. CONCLUSION

In this letter, we proposed a novel skip attention (SA) mechanism for MSVS. The proposed SA implements an attention-driven and dependency modeling for the repetitive structures of music sources. In addition, it effectively controls the flow of the low-level (vocal and musical) features in the encoder-decoder structure, which not only can improve the bottleneck structure to extract more essential vocal or musical features but also retrieve those fine and low-level details for high-quality source estimation. Experimental results have shown the effectiveness of the proposed SA-SHN on two different datasets.

References

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

[2] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] I. J. Goodfellow, Y. Bengio, and A. C. Courville, *Deep Learning* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2016.

[4] A. J. R. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," in *Proc. 12th Int. Conf. Latent Variable Anal. Signal Separation*, 2015, pp. 429–436.

[5] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. 15th Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 477–482.

[6] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *Proc. 13th Int. Conf. Latent Variable Anal. Signal Separation*, 2017, pp. 258–266.

[7] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2017, pp. 1265–1269.

[8] J. Liu and Y. Yang, "Denoising auto-encoder with recurrent skip connections and residual regression for music source separation," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl.*, 2018, pp. 773–778.

[9] J. Liu and Y. Yang, "Dilated convolution with dilated GRU for music source separation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 4718–4724.

[10] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-net convolutional networks," in *Proc. 18th Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 745–751.

[11] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," in *Proc. 19th Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 289–296.

[12] C. A. Huang *et al.*, "An improved relative self-attention mechanism for transformer with application to music generation," *CoRR*, abs/1809.04281, 2018.

[13] J. Paulus, M. Müller, and A. Klapuri, "State of the art report: Audio-based music structure analysis," in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 625–636.

[14] 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 Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 8, pp. 1307–1335, Aug. 2018.

[15] H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[16] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[17] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, abs/1609.08144, 2016.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Rep.*, 2015.

[19] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1243–1252.

[20] Y. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *Proc. 14th Int. Soc. Music Inf. Retrieval Conf.*, 2013, pp. 427–432.

[21] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[22] J. Sebastian and H. A. Murthy, "Group delay based music source separation using deep recurrent neural networks," in *Proc. Int. Conf. Signal Process. Commun.*, 2016, pp. 1–5.

[23] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, abs/1607.06450, 2016.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[25] C. Hsu and J. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 310–319, Feb. 2010.

[26] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Proc. 12th Int. Conf. Latent Variable Anal. Signal Separation*, 2015, pp. 387–395.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Rep.*, 2015.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[29] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[30] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 334–340.

[31] J. L. Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 66–70.

[32] I. Jeong and K. Lee, "Singing voice separation using RPCA with weighted l_1 -norm," in *Proc. 13th Int. Conf. Latent Variable Anal. Signal Separation*, 2017, pp. 553–562.

[33] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel music separation with deep neural networks," in *Proc. 24th Eur. Signal Process. Conf.*, 2016, pp. 1748–1752.

[34] S. Uhlich *et al.*, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 261–265.

[35] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 21–25.

[36] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 865–869.