IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 2, 2020, accepted November 9, 2020, date of publication November 16, 2020,
date of current version November 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037773

Parallel Stacked Hourglass Network for Music Source Separation

BHUWAN BHATTARAI, YAGYA RAJ PANDEYA, AND JOONWHOAN LEE
Division of Computer Science and Engineering, Jeonbuk National University, Jeonju 54896, South Korea

Corresponding author: Joonwhoan Lee (chlee@jbnu.ac.kr)

This work was supported in part by the research funds of Jeonbuk National University, in 2020, and in part by the Human Resources Program in Energy Technology of the Korea Institute of Energy Technology Evaluation and Planning (KETEP), granted financial resource from the Ministry of Trade, Industry and Energy, Republic of Korea, under Grant 20204010600470.

ABSTRACT Music source separation is one of the old and challenging problems in music information retrieval society. Improvements in deep learning lead to big progress in decomposing music into its constitutive components with a variety of music. This research uses three types of datasets for source separation namely; Korean traditional music *Pansori* dataset, MIR-1K dataset, and DSD100 dataset. DSD100 dataset includes multiple sound sources and other two datasets has relatively smaller number of sound sources. We synthetically constructed a novel dataset for *Pansori* music and trained a novel parallel stacked hourglass network (PSHN) with multiple band spectrograms. In comparison with past study, proposed architecture performs the best results in real-world test samples of *Pansori* music of any length. The network performance was also tested for the public DSD100 and MIR-1K dataset for strength comparison in multiple source data and found comparable quantitative and qualitative outcomes. System performance is evaluated using median value of signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR) measured in decibels (dB) and visual comparison of prediction results with ground truth. We report better performance in the *Pansori* dataset and MIR-1K dataset and perform detailed ablation studies based on architecture variation. The proposed system is better applicable for separating the music source with voices and single or fewer musical instruments.

INDEX TERMS Music source separation, parallel stacked hourglass network, multiband spectrogram, *Pansori*.

I. INTRODUCTION

Music is a mingling of several signals to form one combined signal. The goal of music source separation is to isolate original music components from the combined signals to better understand music theory.

Music source separation has several useful applications including automatic speech recognition for bilateral cochlear implant patients [1], fundamental frequency estimation for music transcription [2], beat tracking despite the presence of highly predominant vocals [3], the generation of karaoke music, instrument detection, lyrics recognition and chord estimation. Another application is singer identification in the music management system by separating the singing voice from music accompaniment. The source separation also can be useful for education propose, for example teaching the way of playing or singing any rare/traditional song or musical instruments. This research focus to isolate sound sources of Korean traditional music called *Pansori*, one of the intangible heritages registered in UNESCO. The separation is helpful for traditional music automatic transcription or education for preserving. Most of the traditional music including western and Asian music is similar to nature with fewer music instrument sources. Therefore, this study can be beneficial for most of the transcription systems or traditional music learners.

Music separation has a long history of scientific activity as it is known to be a very challenging problem. The data-driven machine learning approach for music source separation has been of great interest to researchers in recent years, and most of the studies have been conducted on western music [4]–[6] and a study [7] is intended in the application of music and speech isolation. To the best of our knowledge, there is no such study for traditional music with few sources. Another problem in this area is the scarcity of labeled datasets. This

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

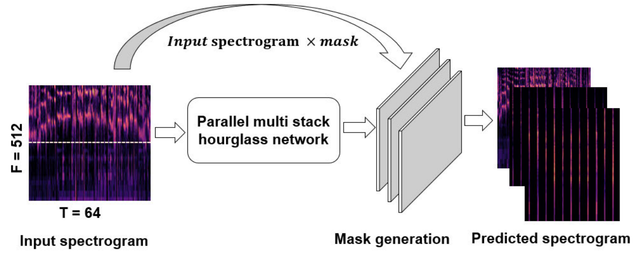

Input spectrogram × *mask*
F = 512
T = 64
Parallel multi stack hourglass network
Input spectrogram
Mask generation
Predicted spectrogram

FIGURE 1. Overview of our source separation.

research focuses on solving these two problems by constructing a new dataset using Korean traditional music, called *Pansori* dataset and introducing a novel network architecture.

A basic overview of our proposed source separation method is shown in Figure 1, where the result of separation is masked to extract music sources. In the method, a parallel stacked hourglass network (PSHN) is proposed for the separation inspired by [8]. The PSHN is a fully convolutional multi-scale network that learns the features of a multiband spectrogram to generate a time-frequency masks. Multiband spectrogram divides the input based on frequency ranges (low to high). This procedure is helpful to isolate musical sources masked by other frequencies. The PSHN was trained and tested on the *Pansori* dataset and publicly available DSD100 and MIR-1K dataset. The main contribution of this study can be summarized as:

1) A new dataset is introduced using Korean traditional music *"Pansori"* with fewer numbers of instrument sounds.
2) A Parallel stacked hourglass network is introduced and tested its capability on various music. The experiment shows that the proposed architecture is beneficial for music with the fewer number of instrument sounds like traditional music or classical music. We compared the proposed method for the various datasets with multiband time-frequency representation. Comparison is made based on the number of hourglasses stacked in network architecture, the number of musical sources in the datasets, and the robustness of the proposed system than the previous work.

The rest of this paper is organized as follows. Section II describes the related work, and some specific results obtained by other researchers seeking to improve the quality of audio source separation. Section III introduces the *Pansori* dataset and its construction. The proposed PSHN is provided in Section IV. Section V provides detailed results of our experiments, comparison with previous research and detailed ablation studies based on the architectural variation of PSHN. Finally, in Section VI we offer our concluding thoughts.

II. RELATED WORK

The past decade has witnessed the creation of several new approaches to the problem of separating individual audio sources from the mixture. Independent Component Analysis (ICA) [9], non-negative matrix factorization (NMF) [10], and sparse component analysis (SCA) [11] are some of the statistical techniques that have been developed for blind Source Separation [12], [13]. The essential thought behind these methods is to project the data from a time series into a new set of axes that depend on some statistical approach.

More recently, deep learning strategies have demonstrated their superiority over these methods and allowed for a significant improvement in source separation. In general, neural networks adopt a hierarchical architecture to read latent information from the audio data. It can be developed as a non-linear approximation function to estimate the independent music source from the combined signals. In [14], a fully connected network was used to separate the spectra of the source using 2D Mel-Spectrogram. A fully convolutional denoising auto-encoder (CDAE) was proposed in [15] to do single-channel source separation. Their goal was to see if the CDAE could learn the spectral-temporal filters and features to its corresponding source. In [16] a deep neural network (DNN) with five fully connected layers and ReLU nodes was designed for instrumental source separation. Similarly, the work in [17] also trained a DNN for source separation using multichannel audio input by focusing entirely on spectral characteristics of a single frame.

Additionally, recurrent neural networks (RNN) were used to separate singing voices in single-channel musical recordings in [18]–[21] to preserve temporal information. The most accurate method so far has been that applied by [22], which involved blending a feed-forward neural network and bidirectional long-short term memory (BLSTM) approach. This two-network blending technique suggests that even where the network structure and training method are different, combining the two is beneficial for music source separation purposes.

Some studies use waveform music representation for source separation to preserve the phase information of the audio signal. The work in [23] used U-Net based architecture, which resamples the features at different time scales. A modified U-Net architecture improved performance by including source additivity in the output layer, upsampling, and employing context-aware prediction. Similarly, [24], and [25] used an encoder-decoder model that addressed the source separation problems of multiple speakers with multiple audio channels. The waveform representation is also applied for isolation of speech from noisy signals [26]–[28]. The major challenges using waveform audio representation is to preserve sinusoid audio information [29] and management of memory required using a data driven-approach.

Multiband time-frequency audio representation is popular for source separation and achieved state-of-the-art results on the DSD100 dataset. The idea behind this representation is a division of spectrogram based on multiple frequency bands as the spectrogram contains high energies in the lower band and low energies in the higher band. Applying different convolution in the spectrogram band containing higher and lower energies, the network could efficiently model both global and local features. Researches in [4], [5], and [7] adopted DenseNet architecture [30] for music source separation using multiband spectrogram. This type of representation is also found effective for sound classification [31]. Motivated by these past studies, we split the spectrogram into multiple sub-bands along the frequency axis and the sub-bands and full band were passed into our PSHN. This representation was found effective for this *Pansori* source separation.

In conclusion to these past research, the music source separation is limited to little dataset and music variety. To boost this research area, first, we construct a labeled dataset for source separation of Korean traditional music. The length of sources for drum and drummer voice is very short and consistent throughout the whole music which challenges the music source separation community to separate the short time sources of music. Similarly, the source for singer's voice has a different frequency range as both the storytelling part and singing part is performed by single artist. This makes the *Pansori* dataset unique and challenging over the current music source separation dataset. The dataset is synthetic but it well preserves the real-world music pattern and hence our network performance does not bias in real test samples. The network architecture is data-dependent and it should be compared with multiple dataset. So, to validate the network architecture, the two publicly available DSD100 and MIR-1K datasets are compared to the past state-of-the-art network. PSHN achieves the best performance for the MIR-1K dataset and singing voice separation task of the DSD100 dataset. These two experiments only contain two sources singing voice and accompaniments. The only two sources on MIR-1K and singing voice separation task of DSD100 dataset achieves better results than the multiple music source separation task of the DSD100 dataset. Therefore, the statement clarifies that the proposed PSHN is better for music with a fewer number of sources.

III. PANSORI DATASET

Pansori music is emerged in South Korea in the late seventeenth century and has become popular amongst the privileged class by the middle of the eighteenth century. Now, it has been registered as intangible heritage by UNESCO, which should be preserved and transferred to the next generation in the future. *Pansori* music consists of two performers; a singer and a drummer. The drummer keeps time and provides appropriate rhythmic accompaniment taking cues from the singer, while the singer tells a melodic story and takes cues from the drummer. *Pansori* is divided into two portions: a storytelling time wherein the singer explains various characters and expresses the feelings of a story. In the second part, the singer projects his or her voice from their stomach. The power thus generated resonates with the audience in a powerful way. Therefore, there are fundamentally three sources in *Pansori* music; drum, drummer voice, and singer's voice. The drum and drummer voice are repeated throughout the song, but both are played in a short timeframe; the drummer's voice is under one second while the drum may last from one to three seconds. Drum sounds may be one of three distinct kinds or a blend of these. While a drummer's primary job is to energize the vocalist and account for the vocalist's physical rhythm to elicit the best possible performance from the singer, their second job is to carefully observe the singer's mouth to anticipate their next breath or sound.

TABLE 1. Data statistics of Korean traditional *Pansori* song.

Data	Drum	Drummer voice	Singer voice	Mixture
Training data	50	50	50	50
Test data	15	15	15	15

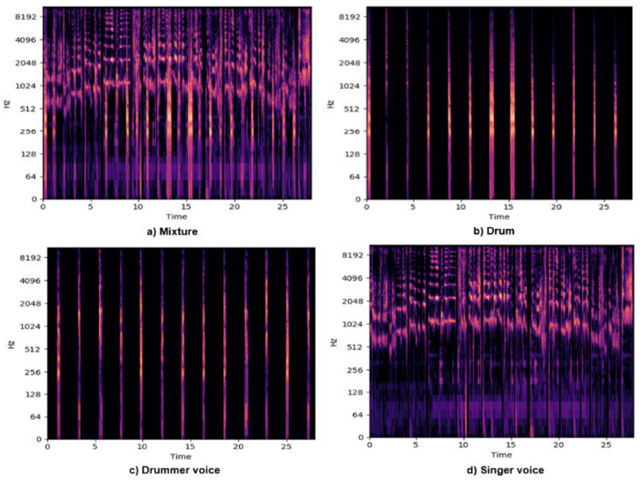
Pansori is a Korean genre of musical storytelling performed by a singer and a drummer. Generally, *Pansori* is performed by a female singer with a male drummer. So, our *Pansori* dataset included only the female singer with male drummer. The drummer also produces sound during the singer's performance which we denote as the drummer's voice source. To achieve source separation in *Pansori*, a dataset was constructed from numerous performers from YouTube videos and compact disc (CD) recordings. The YouTube videos are recorded during the stage performance of the artist and music on CD is studio recorded. Even the music is recorded in stage performance, there is no much noise in the audio. The music is recorded at the sampling rate of 44.1kHz and the number of microphone channels is equal to 2.

Ground truth of each *Pansori* source is needed to separate them from the mixture. We obtained this by physically removing the drum and drummer voice from the original song and saved it separately in our singer-voice track. The resulting short sample of the drum and drummer voice was cut and saved in such a way that it contained as little singer voice as possible. Samples of drum and drummer voice were removed if the signal for the singer's voice appeared on it. After the removal of the drum and drummer voice, the signal was saved as the ground truth for the singer's voice. For making the ground-truth signal for the drum and drummer voice, we built a silence signal of equivalent length to the ground truth singer voice signal and embedded the short clip of drum and drummer voice into the silent for an arbitrary span. Then blended these three ground truth sources, resulting in mixed sources. In total, fifty training audio data files were created for each source and mixture, while fifteen files of test audio data were created. The duration of each data samples is in the varying length of ranged from five-seconds to five minutes, with an average duration of fifty-one seconds. Statistics related to this *Pansori* dataset are shown in Table 1. The log spectrogram visualization of three ground truth sources along with their mixture is shown in Figure 2. The *Pansori* dataset and experimental results are publicly available on GitHub.[1]

IV. METHOD

This section first reviews the basics of the stacked hourglass network (SHN). It then extends an SHN to our proposed parallel stacked hourglass network (PSHN) with intermediate

[1] https://github.com/pratikshaya/pansori_source_separation



a) Mixture

b) Drum

c) Drummer voice

d) Singer voice

FIGURE 2. Log spectrogram visualization of *Pansori* (a) Mixture, (b) Drum, (c) Drummer voice and d) Singer voice.

predictions. Finally, we provide details about the loss function for training PSHN for music source separation.

A. STACKED HOURGLASS NETWORK

Hourglass modules (HMs) were first introduced as a means of estimating human poses in color images [1]. They are fully convolutional networks which extract the multi-scale features using a top-down and bottom-up approach. The multi-scale features of HMs capture the relevant information at various scales and resolutions. The input spectrogram is first passing into the series of five initial convolutions. The first convolution layer has a filter size of 7*7 to obtain the larger receptive field from the input spectrogram followed by four 3*3 convolutional layers. The output feature map for each layer is 64, 128, 128, 128, and 256 respectively. As applied to music sources, local evidence is essential to map the short period relationship while the global context is essential to determining the long-term dependency. Coarse resolution feature maps were obtained through a series of convolution and pooling operations. Next, the coarse resolution feature maps are upsampled using nearest-neighbor interpolation. The original HM had four downsampling and upsampling operations, but this paper used HM with five downsampling and upsampling operations to obtain the fine resolution feature map.

HMs has a symmetric topology and are stacked together to form a stacked hourglass network (SHN). In this way, the HMs will produce a mask for prediction, with subsequent HMs refining their mask after learning the predictions from the earlier HMs. To demonstrate this, in the experiment section we compared the ground truth spectrograms with the predicted spectrograms and shows the measures using the evaluation metrics. The architecture of a single HM is shown in Figure 3.

Our hourglass module is similar to the one used in [6], which relied on a light-weight version that replaced the residual blocks [32] with a single convolutional layer. This light-weight architecture was competitive with the original stacked hourglass network, with a much smaller number of parameters [33]. Our HM is different only in that we added one more upsampling and downsampling step to obtain a finer resolution.

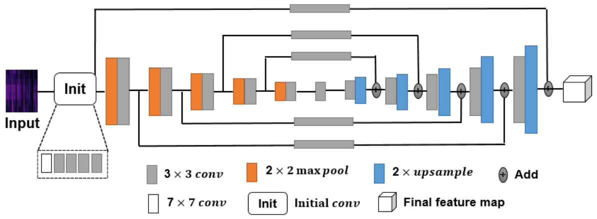


FIGURE 3. Overview of our hourglass module.

B. PSHN AND FEATURE CONCATENTATION FROM MULTI-BAND SPECTROGRAM

A multiband spectrogram is used as input to PSHN. As discussed earlier, spectrograms represent similar patterns in different frequency bands. Especially, low frequencies are bounded by high energies, while, higher frequencies contain low energies and noise. In consequence, the spectrogram is divided into equal halves along the frequency axis to form two sub-band spectrograms and a different convolution filter is applied in each band.

Consider the mixed magnitude spectrogram X of the size $F \times T \times 1$ where F represents the frequency axis, T represent the time axis, and 1 is the spectral channel. We divided the full spectrogram into two halves, and labeled one the upper band spectrogram and another the lower band spectrogram. First, the initial convolutions are applied for each of the bands separately and passed into the PSHN. The size of each divided band was equal to $F/_2 \times T \times 1$.

The parts of the stacked hourglass network (SHN) that receive the input from the upper band, lower band, and full band, and those are passing through an upper band stacked hourglass network (UBSHN), lower band stacked hourglass network (LBSHN), and full band stacked hourglass network (FBSHN), respectively. The combination of all three is a PSHN. Each of the UBSHN, LBSHN, and FBSHN contains four stacked hourglass modules in total. The hourglass modules of UBSHN and LBSHN output feature maps of size $F/_2 \times T \times C$, where, in our case, $C = 256$ is the channel number. To obtain the full feature map of size $F \times T \times C$ in all stack of hourglass modules, we concatenated the output feature map in the frequency axis of UBSHN and LBSHN. Likewise, each hourglass module from the FBSHN output a feature map of size $F \times T \times C$. The full feature map from UBSHN and LBSHN is then again concatenated, now with the feature map obtained from FBSHN, to get the final feature map of size $F \times T \times 2C$. After getting the feature map of size $2C$, two consecutive 1×1 convolutions are applied to reduce the dimensions of the feature map and estimate the masks of each music source. The prediction of each mask is multiplied with the input to obtain the predicted

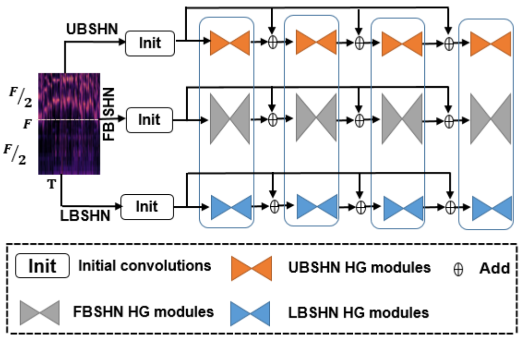


FIGURE 4. Proposed parallel stacked hourglass network for multi band spectrogram.

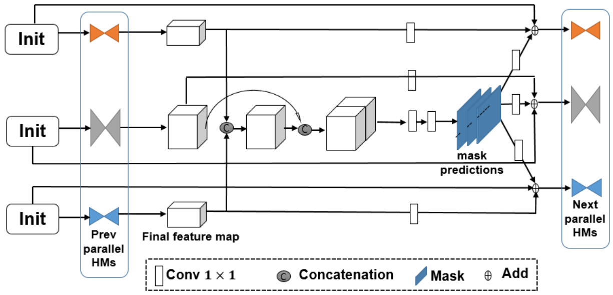


FIGURE 5. Intermediate predictions between previous HMs and the subsequent HMs.

spectrogram. The loss between the predicted spectrogram and the ground truth spectrogram was then calculated. The proposed architecture for PSHN for a multiband spectrogram is shown in Figure 4.

C. INTERMEDIATE PREDICTIONS

The process of estimating the masks and predicting the spectrograms between an earlier and a subsequent parallel hourglass module is referred to as intermediate predictions. There are four stacks of hourglass modules for the upper band, lower band, and full band spectrograms. Accordingly, we needed to calculate four losses: three for the intermediate predictions and one for the final prediction. All predictions result predicts the full-size spectrogram of size $F \times T \times S$, from the full mask (FM) where S represents the number of sources to separate. Recall also that the FM is divided into an upper mask (UM) and a lower mask (LM). The 1×1 convoluted UM final feature map obtained from the previous HM and the input of the previous hourglass module is added together and passed to the next hourglass module of the UBSHN. This operation is repeated for the HM of the FBSHN and the LBSHN, with FBSHN receiving the FM and the LBSHN receiving the LM. Prior to adding the mask feature for the next hourglass module, 1×1 convolution is applied to make the feature dimension equal to that of the final feature map obtained from the previous HM. The intermediate predictions step, which takes place between the previous and next parallel HMs is shown in Figure 5.

D. LOSS FUNCTION

The objective of our study was to isolate the three independent sources of *Pansori*. First, the magnitude spectrogram of the mixed source was divided along the frequency axis and passed into the PSHN. As *Pansori* contains three sources, the size of the output masks in each prediction of PSHN was $F \times T \times 3$. The generated masks were multiplied separately with the ground truth spectrograms of the singer's voice, drummer voice, and drum to generate the predicted spectrograms. The experimental result shows that stacking parallel HMs produced more refined masks of estimated spectrograms from each prediction.

The loss function used in this work is the $L_{1,1}$ norm, which is used to minimize the absolute difference between the target spectrogram and the estimated spectrogram. The FM output of PSHN for the i^{th} source spectrogram in j^{th} prediction is represented as M_{ij}. The input spectrogram X is multiplied with M_{ij} to obtain the estimated spectrogram. The PSHN is trained to minimize the absolute difference between the i^{th} target spectrogram of the music source Y_i and the generated spectrogram. The $L_{1,1}$ norm loss function for the i^{th} source spectrogram in the j^{th} prediction is defined as:

$$Loss\,(i, j) = \parallel Y_i - X \odot M_{ij} \parallel_{1,1} \tag{1}$$

where $\odot$ represents an element-wise matrix multiplication and $\parallel . \parallel_{1,1}$ represents 1-norm, which is the sum of absolute values of each element of the matrix. There are four hourglass modules stacked together in each parallel network of the PSHN architecture. So, the total loss of the network is calculated by summing up these four predictions.

$$Totalloss = \sum_{i=1}^{S} \sum_{j=1}^{4} Loss\,(i, j) \tag{2}$$

Here, S represent the number of sources to separate. So, for the *Pansori* dataset $S = 3$, for music source separation task of DSD100 dataset $S = 4$, and for singing voice separation task of DSD100 dataset and MIR-1K dataset $S = 2$.

V. EXPERIMENTS

We performed our first PSHN experiment on our *Pansori* dataset. We evaluated the performance of the system by using the median of Signal-to-distortion ratio (SDR) value, source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR) measured in decibels (dB), and based on BSS-Eval metrics [36]. We then report the performance in two publicly available DSD100 and MIR-1K datasets. DSD100 dataset was prepared for SiSEC 2016 [37]. The median (SDR) value is calculated from *Pansori* and DSD100 datasets. Whereas, global normalized SDR (GNSDR), global SIR (GSIR), and global SAR (GSAR) is calculated as a weighted mean of NSDR, SIR, and SAR respectively by following [6] for the MIR-1K dataset. All three datasets test results are measured from the PSHN

(1-Stack), PSHN (2-Stack), PSHN (3-Stack), and finally from the PSHN (4-Stack).

A. EXPERIMENTAL CONFIGURATION

The audio file from the datasets was preprocessed with the Librosa library [34], which we used to generate the magnitude spectrogram at a sampling rate of 8kHz. The duration of our music sources are varied in length for the *Pansori*, MIR-1K and DSD100 datasets. From these varying lengths of audio, magnitude spectrograms were calculated with a window size of 1024 and a hop length of 256. The average spectrogram size from the two datasets was 512 × 1593 (*Pansori*) and 512×7812 (DSD100). The average length of MIR-1K dataset is not given. So, it is not possible to calculate the average sized spectrogram. To make a fixed size input for the network, we split longer segments into series of 64 (spectral time) by 512 (frequency bins) fixed-length sub segments that inherit their respective spectral representation for training [6]. This way, we retain more training data. Sub segments are non-overlapping, except for the last one which might overlap with the penultimate when it does not match the network input size. Accordingly, the number of spectrogram samples obtained from the average-sized of spectrograms was equaled to 25 (*Pansori*) and 123 (DSD100).

The training was conducted with Adam optimizer [35] with an initial learning rate of 0.0001. The network was trained through 200000 iterations for all *Pansori*, MIR-1K, and DSD100 datasets, and the learning rate was decreased to 0.00001 when 75% of the training was complete. The GPU used to train the network was an NVIDIA TITAN XP. During testing, we visualized the predicted spectrograms from each of the four predictions of *Pansori* and DSD100 datasets and report the performance from all four predictions in Section V.

B. RESULTS FROM PANSORI DATASET

The *Pansori* dataset had three ground truth sources; drum, drummer voice, and singer's voice. The experiment was performed by single-channel source separation and the median SDR value from the test data is reported from four predictions of PSHN. The median SDR performance of the system from each of the predictions is shown in Table 2.

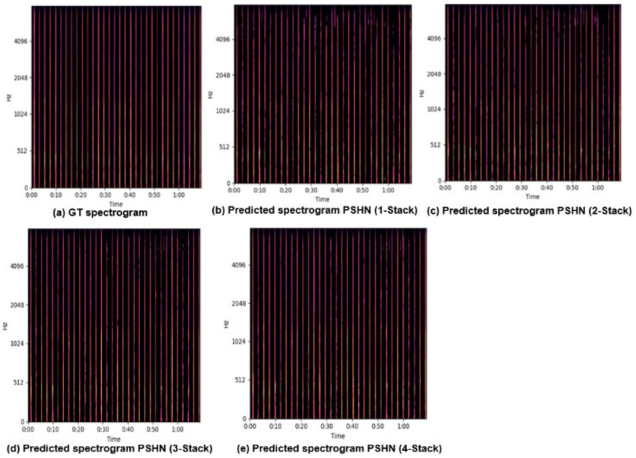
TABLE 2. Median SDR values for *Pansori* source separation dataset.

Method	Drum	Drummer voice	Singer voice
PSHN (1-Stack)	15.65	12.40	15.76
PSHN (2-Stack)	15.81	12.54	15.94
PSHN (3-Stack)	15.89	12.66	16.03
PSHN (4-Stack)	**15.97**	**12.86**	**16.12**

Bold number indicate the highest evaluation score.

All four predictions of the PSHN have better separation results in the *Pansori* dataset. There was a 0.32dB, 0.46dB, and 0.36dB SDR difference between the 1-Stack and the 4-Stack PSHN for the drum, drummer voice, and singer's voice respectively. The final predictions PSHN (4-Stack) performed well for all three sources. The SDR values from these predictions reached as high as 15.97dB, 12.86dB, and 16.12dB respectively for drum, drummer voice, and singer's voice. The length of the drummer's voice is less than one second. These short and unique sounds sometimes do not become part of the training set. Therefore, the SDR for the drummer's voice limits the performance in comparison with drum and singer voice.

Figure 6 shows the qualitative result for one of the audio samples of the drummer's voice in our test set. The vertical line in the log spectrogram represents the drummer voice over a short period of time. The PSHN can accurately separate the sources of *Pansori*, which can be seen by observing the ground truth (GT) log spectrogram and predicted log spectrogram in Figure 6. It is particularly noteworthy that the log spectrograms predicted by PSHN (3-Stack) and PSHN (4-Stack) captured fine details and recovered more at the frequency range above 4096 Hz. The sources for drum and drummer voice were short and consistent across all samples, which helped the network to learn and predict the target sources accurately.



(a) GT spectrogram

(b) Predicted spectrogram PSHN (1-Stack)

(c) Predicted spectrogram PSHN (2-Stack)

(d) Predicted spectrogram PSHN (3-Stack)

(e) Predicted spectrogram PSHN (4-Stack)

FIGURE 6. Results of comparison between ground truth and predicted spectrograms for one drummer voice audio sample in our test set. (a) GT spectrogram, (b) Predicted spectrogram PSHN (1-Stack), (c) Predicted spectrogram PSHN (2-Stack), (d) Predicted spectrogram PSHN (3-Stack), (e) Predicted spectrogram PSHN (4-Stack).

C. RESULTS FROM DSD100 DATASET

The DSD100 (demixing secret database) dataset initially introduced by SiSEC in 2015 [41] consists of 100 full-track songs of different musical styles and genres and is divided into development and test subsets. Later on, the DSD100 dataset, prepared for SiSEC 2016 [37], improved the sound quality, so that for each track, it consists of four semi-professionally engineered stereo source. The dataset contains the music tracks of various styles, along with their isolated bass, drum, vocal, and other sources. The duration of the songs ranges from 2 minutes and 22 seconds to 7 minutes and 20 seconds, with an average duration of 4 minutes and 10 seconds. The four sources are added together to form mixed sources. The dataset is divided into fifty training sets

TABLE 3. Median SDR values for music source separation for DSD100 dataset.

Method	Bass	Drums	Other	Vocals
PSHN (1-Stack)	1.85	4.36	2.39	5.40
PSHN (2-Stack)	2.27	4.48	2.49	5.59
PSHN (3-Stack)	2.32	4.49	2.52	5.60
PSHN (4-Stack)	**2.35**	**4.52**	**2.55**	**5.70**
SH-4stack [6]	1.77	4.11	2.36	5.16

Bold number indicate the highest evaluation score.

TABLE 4. Median SDR values for singing voice separation for DSD100 dataset.

Method	Vocals	Accompaniments
PSHN (1-Stack)	5.26	11.87
PSHN (2-Stack)	5.76	12.27
PSHN (3-Stack)	5.87	12.38
PSHN (4-Stack)	**6.01**	**12.42**
SH-4stack [6]	5.45	12.14

Bold number indicate the highest evaluation score.

and fifty test sets. All signals are stereophonic and encoded at a sampling rate of 44.1kHz. The experiment was performed by single-channel source separation and the median of SDR value from the test data was reported from each prediction of PSHN.

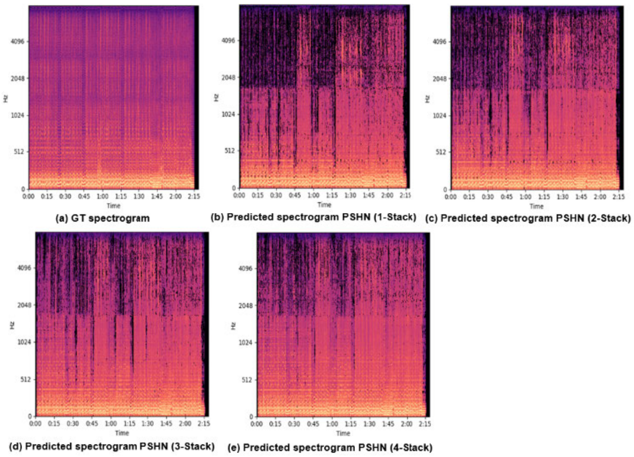
The task for the DSD100 dataset was twofold: (1) music source separation for a sample with four independent sources of bass, drums, other, and vocals, and (2) singing voice separation for a sample with two independent sources of vocals and accompaniments.

The performance (median SDR value) of PSHN and baseline SH-4stack [6] at music source separation from the sample with four different independent sources is shown in Table 3. Our tests revealed that SDR value increases as HMs across parallel hourglass network are stacked together. Among all four predictions, PSHN (4-Stack) had the best source separation results, with bass seeing the greatest difference in SDR between the PSHN (1-Stack) and the PSHN (4-Stack) (a difference of 0.50dB). The drum and other categories saw an increase in SDR of 0.16dB between PSHN (1-Stack) and PSHN (4-Stack), while, Vocals also increased by 0.30dB. The smallest difference in SDR is between the PSHN (2-Stack) and the PSHN (3-Stack), which was only 0.05dB for bass, 0.01dB for drum, 0.03dB for other, and 0.10dB for a vocal.

We also used the DSD100 dataset to determine SDR values associated with a singing voice separation task. To perform this experiment, three sources of DSD100 (excluding vocals) are blended together to obtain an accompaniment source. The result was that a PSHN architecture trained with the sources of vocals and accompaniments saw improvements in its performance at vocal separation. The performance of the PSHN and baseline SH-4stack [6] during the singing voice separation experiment is shown in Table 4. It should be noted that the PSHN (4-stack) in this task achieved even better result which is 0.31dB more in comparison with the music source separation task.

The change in performance with vocals in two different tasks of DSD100 illustrates that our PSHN architecture improved as the number of sources to separate decreased. This finding indicates that the hourglass configuration is superior for identifying three sources of *Pansori*.

The qualitative results from one of the test set of audio examples in the DSD100 dataset for the source bass are shown in Figure 7. The ground truth and PSHN predicted log spectrograms from all predictions are provided. The most remarkable difference observed between the predicted spectrogram of PSHN (1-Stack) and PSHN (4-Stack). As the mask was refined after stacking HMs in the parallel stacked network, succeeding HMs were able to capture fine details of formants and harmonics from the predicted spectrogram.



(a) GT spectrogram (b) Predicted spectrogram PSHN (1-Stack) (c) Predicted spectrogram PSHN (2-Stack)

(d) Predicted spectrogram PSHN (3-Stack) (e) Predicted spectrogram PSHN (4-Stack)

FIGURE 7. Results of comparison between ground truth and predicted spectrograms for one bass audio sample from the DSD100 dataset. (a) GT spectrogram, (b) Predicted spectrogram PSHN (1-Stack), (c) Predicted spectrogram PSHN (2-Stack), (d) Predicted spectrogram PSHN (3-Stack), (e) Predicted spectrogram PSHN (4-Stack).

D. RESULTS FROM MIR-1K DATASET

The MIR-1K dataset is designed for singing voice separation tasks. It contains 1000 song clips in which the musical accompaniment and the singing voice are recorded at the left and right channels, respectively at a sampling rate of 16kHz. Manual annotation of the dataset includes pitch contours in semitone, indices, and types for unvoiced frames, lyrics, and vocal/non-vocal segment. The song clip is named in the form *"SingerId_SongId_ClipId.* The duration of each clip ranges from 4 to 13 seconds, and the total length of the dataset is 133 minutes. These clips are extracted from 110 Karaoke songs which contain a mixture track and a music accompaniment track. The songs are sung by 8 females and 11 males and most of the singers are amateur and do not have professional music training. Following the work [6], we used one male and one female (*abjones and amy*) as a training set which contains 175 clips in total. The remaining 825 clips are used for testing.

The performance (GNSDR, GSIR, and GSAR) of PSHN and baseline SH-4stack [6] at singing voice separation from

TABLE 5. GNSDR, GSIR, and GSAR values for singing voice separation on MIR-1K dataset.

Singing voice			
Method	**GNSDR**	**GSIR**	**GSAR**
PSHN (1-Stack)	10.35	15.94	12.30
PSHN (2-Stack)	10.66	16.22	12.61
PSHN (3-Stack)	10.81	16.54	12.67
PSHN (4-Stack)	**10.83**	**16.54**	**12.67**
SH-4stack [6]	10.51	16.01	12.53

Accompaniments			
Method	**GNSDR**	**GSIR**	**GSAR**
PSHN (1-Stack)	9.54	13.67	12.33
PSHN (2-Stack)	9.77	13.88	12.56
PSHN (3-Stack)	9.83	13.92	12.64
PSHN (4-Stack)	**9.89**	14.01	**12.65**
SH-4stack [6]	9.88	**14.24**	12.36

Bold number indicate the highest evaluation score.

the sample with two sources singing voice and accompaniments is shown in Table 5. As in the DSD100 dataset, the test results on the MIR-1K dataset also revealed that the performance increases as HMs across parallel hourglass networks are stacked together. It should be noted that, among all four predictions, PSHN (4-Stack) had the best separation performance in all evaluation criteria except GSIR of accompaniments. PSHN (4-Stack) in singing voice gains 0.48dB in GNSDR, 0.60dB in GSIR, and 0.37dB in GSAR compared to the results of PSHN (1-Stack). Whereas, it gains 0.35dB in GNSDR, 0.34dB in GSIR, and 0.32dB in GSAR for the accompaniments.

E. BASELINE COMPARISON

We compared our PSHN architecture with the baseline architecture of SH-4stack [6] in MIR-1K dataset and both music source separation and singing voice separation tasks of the DSD100 dataset. The baseline, which uses a single full band spectrogram and no parallel hourglass modules performed worse than the PSHN. While the same convolutional kernel was applied to the baseline architecture of SH-4stack, the PSHN, which used a different convolutional kernel for the multiband spectrogram, tended to improve the performance. The results obtained from the first test of our PSHN at music source separation on DSD100 dataset are provided in Table 3. The PSHN (1-Stack) performed better than even the fourth stacked of the baseline SH-4stack. The baseline achieved 1.77dB, 4.11dB, 2.36dB, and 5.16dB SDR for bass, drums, other, and vocals respectively. These SDR values are less than all four predictions results obtained from the PSHN. The PSHN (1-Stack) achieved 1.85dB and the PSHN (4-Stack) achieved 2.35dB for bass, which are 0.08dB and 0.58dB more than the SH-4stack. Similarly, PSHN (4-Stack) showed its superior performance in identifying the sources of drums, other, and vocals, with drums achieving 0.41dB, other achieving 0.19dB, and vocals achieving 0.58dB gains in comparison with the baseline.

The task for the DSD100 dataset can be broken into twofold. One is for music source separation task with four independent sources and another is for singing voice separation task with two independent sources of vocal and accompaniments. For the task of singing voice separation, three sources of DSD100 except vocals are mixed to get an accompaniment source. A comparison between the PSHN architecture and the baseline SH-4stack at singing voice separation task using the DSD100 dataset are provided in Table 4. The PSHN architecture demonstrated that it significantly improved performance for both vocals and accompaniments. This shows the advantage of a multiband spectrogram over the baseline by significantly increasing the SDR values across all predictions of PSHNs. The baseline with no parallel hourglass modules achieved 5.45dB for vocals and 12.14dB for accompaniments. Compare with PSHN (4-stack), the baseline achieved 0.56dB less for vocals and 0.28dB less for accompaniments. The PSHN (2-stack) and PSHN (3-stack) also outperformed the baseline results, showing increased SDR for vocals by 0.31dB and 0.42dB and for accompaniments by 0.13dB and 0.24dB respectively.

Similarly, the results obtained from the MIR-1K dataset of our PSHN and baseline SH-4stack [6] at singing voice separation are provided in Table 5. PSHN (2-Stack), PSHN (3-Stack), and PSHN (4-Stack) achieve better results than the baseline while, PSHN (1-Stack) slightly decreases the performance in all evaluation measures in comparison with baseline by 0.16dB in GNSDR, 0.07dB in GSIR, and 0.23dB in GSAR for singing voice. Besides, for accompaniments, the performance by PSHN (4-Stack) is higher than the baseline in GNSDR and GSAR. However, the performance is decreased by 0.23dB in GSIR using PSHN (4-Stack). It also noted that there is less difference in performance between PSHN (3-Stack) and PSHN (4-Stack) in both singing voice and accompaniments. The experimental results on the MIR-1K dataset also support that the multiband spectrogram using a parallel stack hourglass network improves the performance.

F. COMPARISON WITH STATE-OF-THE-ART

The performance of our PSHN (4-stack) in the MIR-1K dataset and both tasks of the DSD100 dataset was compared with previous state-of-the-art method:

1) DeepNMF [38]: This method utilizes non-negative deep network architecture which results from unfolding the NMF iterations by untying its parameter.
2) NUG [14]: This method estimated the source spectra using deep neural networks combined with spatial covariance matrices to encode the source spatial characteristics.
3) BLEND [22]: This approach blended a feed-forward neural network and a recurrent neural network by combining their raw outputs using multi-channel Wiener filtering. We compared single-channel methods of

TABLE 6. Comparison of SDR between various state-of-the-art methods for music source separation for DSD100 dataset.

Method	Bass	Drums	Other	Vocals
DeepNMF [38]	1.88	2.11	2.64	2.75
NUG [14]	2.72	3.89	3.18	4.55
BLEND [22]	2.76	3.93	3.37	5.13
SH-4stack [6]	1.77	4.11	2.36	5.16
MMDenseNet [4]	**3.91**	5.37	3.81	6.00
MMDenseLSTM [5]	3.73	**5.46**	**4.33**	**6.31**
PSHN (4-Stack)	2.35	4.52	2.55	5.70

Bold number indicate the highest evaluation score.

BLEND for the music source separation task and multi-channel methods for the singing voice separation task.

4) MMDenseNet [4]: This method built a parallel DenseNets for full band spectrogram and sub-band spectrogram.
5) MMDenseLSTM [5]: This method enhances MMDenseNet [4] by integrating long short-term memory (LSTM) in multiple scales with skip connections.
6) MLRR [39]: This method uses online dictionary learning to learn the subspaces and propose an algorithm called multiple low-rank representations (MLRR) to decompose a magnitude spectrogram into two low-rank matrices.
7) U-Net [40]: This method adapts the U-net architecture to the task of singing voice separation by comparing the benefits of low-level skip connections with a plain convolutional encoder-decoder model.
8) SH-4stack [6]: This is the baseline of our PSHN method which uses a stacked of four hourglass network.

Tables 6 and 7 and 8 compare our PSHN (4-stack) with the eight other state-of-the-art methods described above. It is proven that our PSHN architecture significantly outperforms the existing methods MLRR [39], U-Net [40], and all stack hourglass networks [6] in all evaluation criteria except GSIR of accompaniments in the MIR-1K dataset. Our method gains by a large margin which is 3.4dB in GNSDR, 4.75dB in GSIR, and 2.25dB in GSAR for singing voice and 2.44dB in GNSDR, 2.58dB in GSIR, and 2.24dB in GSAR for accompaniments in compare with U-Net [40]. Similarly, MLRR which uses the low-rank representation of singing voice and accompaniments also performs worse than our PSHN.

Our PSHN architecture achieved the second-best performance among the current state-of-the-art methods for both vocals and accompaniments at the singing voice separation task (Table 7) on the DSD100 dataset. It was, however, only the third-best performer at the music source separation task (Table 6) of separating drums and vocals, and it showed poor comparative performance at separating bass and other sources. The reason for this is the similarity between bass and other sounds, which confused the PSHN network when it was trained on those sounds together. The sounds of drums and vocals, in contrast, are easier to differentiate from each other.

TABLE 7. Comparison of SDR between various state-of-the-art methods for singing voice separation for DSD100 dataset.

Method	Vocals	Accompaniments
DeepNMF [38]	2.75	8.90
NUG [14]	4.55	10.29
BLEND [22]	5.23	11.70
SH-4stack [6]	5.45	12.14
MMDenseNet [4]	6.00	12.10
MMDenseLSTM [5]	**6.31**	**12.73**
PSHN (4-Stack)	6.01	12.42

Bold number indicate the highest evaluation score.

TABLE 8. Comparison of GNSDR, GSIR, and GSAR between various state-of-the-art methods for singing voice separation for MIR-1K dataset.

Singing voice			
Method	**GNSDR**	**GSIR**	**GSAR**
MLRR [39]	3.85	5.63	10.70
U-Net [40]	7.43	11.79	10.42
SH-1stack [6]	10.29	15.51	12.46
SH-2stack [6]	10.45	15.89	12.49
SH-4stack [6]	10.51	16.01	12.53
PSHN (4-Stack)	**10.83**	**16.54**	**12.67**

Accompaniments			
Method	**GNSDR**	**GSIR**	**GSAR**
MLRR [39]	4.19	7.80	8.22
U-Net [40]	7.45	11.43	10.41
SH-1stack [6]	9.65	13.90	12.27
SH-2stack [6]	9.64	13.69	12.39
SH-4stack [6]	9.88	**14.24**	12.36
PSHN (4-Stack)	**9.89**	14.01	**12.65**

Bold number indicate the highest evaluation score.

So it improves the performance as the losses for all sources in PSHN from all predictions are summed up with equal weights. From this, we can conclude that PSHN architecture is most accurate when the number of sources to separate is small. This is also shown through the superior results that were achieved at the singing voice separation task of DSD100 dataset in Table 7 and MIR-1K dataset in Table 8 which involved the separation of only two sources.

The performance of the MMDenseLSTM is better than our PSHN because it blends two networks MMDenseNet [4] and LSTM. The combination of two networks with differ in network structure and materials is beneficial because the errors of individual system are uncorrelated [5], [22]. Whereas, the proposed PSHN uses a single unified network with a high correlation on losses.

G. ABLATION STUDY

We now analyze the components of our PSHN method and demonstrate its impact. The ablation study is performed on music source separation of the DSD100 dataset, as it is highly popular in the music source separation community. The performance of the system is evaluated by using the median of Signal-to-distortion ratio (SDR) value. The ablation study is performed based on the effect of the number of upsampling and downsampling steps on the hourglass module, with or without intermediate predictions in the loss function, the effect of the number of stacks more or less than 4, and the effect of the number of spectrogram bands. The letter and symbol for the architectural variation in Table 9 are represented as S = number of the stack, U = number of upsampling and downsampling steps in hourglass module, B = number of spectrogram band, WD = without intermediate predictions, W = with intermediate predictions; the symbol $+$, $++$, and $+++$ indicate the network with those symbols have same architecture. All ablation studies are presented in Table 9.

TABLE 9. Ablation study on DSD100 dataset measured by Signal-to-distortion ratio (SDR) value based on the effect of architectural variation on PSHN.

Number of upsampling and downsamping steps				
Method	**Bass**	**Drum**	**Other**	**Vocal**
PSHN_4S_4U_2B	1.70	4.51	2.15	5.03
PSHN_4S_5U_2B^{++}	**2.35**	**4.52**	**2.55**	**5.70**
SH_4S_5U_1B^{+++}	1.91	4.32	2.43	5.35
SH_4S_4U_1B^{+} [6]	1.77	4.11	2.36	5.16
With/Without intermediate predictions				
PSHN_4S_5U_2B_WD	2.03	4.32	2.40	5.45
PSHN_4S_5U_2B_W^{++}	**2.35**	**4.52**	**2.55**	**5.70**
SH_4S_4U_1B_WD	1.80	4.36	2.44	5.56
SH_4S_4U_1B_W^{+} [6]	1.77	4.11	2.36	5.16
The number of spectrogram bands				
SHN_4S_5U_1B^{+++}	1.91	4.32	2.43	5.35
PSHN_4S_5U_2B^{++}	**2.35**	**4.52**	**2.55**	**5.70**
PSHN_4S_5U_4B	2.16	4.26	2.47	5.50
The number of stacks				
PHN_1S_4U_2B	1.81	4.07	2.39	4.95
PSHN_4S_5U_2B^{++}	**2.35**	**4.52**	**2.55**	**5.70**
PSHN_5S_5U_2B	1.96	4.48	2.63	5.50
PSHN_6S_5U_2B	1.89	4.09	2.42	5.56

Bold number indicate the highest evaluation score.

1) THE IMPACT OF THE NUMBER OF UPSAMPLING AND DOWNSAMPLING STEPS

We compared the experiment of our PSHN and SH_4S_4U_1B^{+} [6] method with four and five downsampling and upsampling steps of the hourglass module. To do this, the experiment on the baseline SH_4S_4U_1B^{+} [6] is extended by increasing one more upsampling and downsampling steps which we called it SH_4S_5U_1B^{+++}. The SH_4S_5U_1B^{+++} which receives full band spectrogram of size $F \times T \times 1$ as input decreases the feature map size in the fifth downsampling layer by 50% than that of SH_4S_4U_1B^{+} [6]. The fine resolution feature maps of size ${}^{F}/_{2^{5}} \times {}^{T}/_{2^{5}} \times C$ after the fifth downsampling steps for SH_4S_5U_1B^{+++} tends to increase the performance in comparison with SH_4S_4U_1B^{+} [6]. Similarly, the PSHN_4S_5U_2B^{++} and the PSHN_4S_4U_2B receives frequency division spectrogram of size ${}^{F}/_{2} \times T \times 1$ as input. The fine resolution feature maps of PSHN_4S_5U_2B^{++} after the fifth downsampling steps of size ${}^{F}/_{2} \times 2^{5} \times {}^{T}/_{2^{5}} \times C$ achieve the greater result in comparison with feature maps of PSHN_4S_4U_2B after the fourth downsampling steps of size ${}^{F}/_{2} \times 2^{4} \times {}^{T}/_{2^{4}} \times C$. These experiments indicate that fine resolution feature maps are important than that of coarse resolution feature maps in music source separation task.

2) WITH/WITHOUT INTERMEDIATE PREDICTIONS IN THE LOSS FUNCTION

The spectrograms predicted after estimating the mask in between an earlier and subsequent parallel hourglass modules are referred to as intermediate predictions (see Figure 5). To investigate the impact of with/without intermediate predictions, we next do the two experiments without intermediate predictions in the loss function. One for our PSHN architecture and the other for the baseline [6]. The loss is calculated only from the final stack of the hourglass module. Thus, the architectural variations on loss function only estimate the masks and spectrograms from the final stack of the hourglass module. The experimental results show that PSHN using 2 band spectrograms with intermediate predictions (PSHN_4S_5U_2B_W^{++}) still achieves higher accuracy compare with PSHN without intermediate predictions (PSHN_4S_5U_2B_WD). But the experimental results without intermediate predictions using 1 band (SH_4S_4U_1B_WD) is higher than the baseline for drum, other, and vocal by 0.25dB, 0.08dB, and 0.40dB respectively.

3) THE NUMBER OF SPECTROGRAM BANDS

The effect of spectrogram bands is compared with 1 band, 2 bands, and 4 bands. There is no parallel network for a single band spectrogram as there is no division in the frequency axis for a single band. We called these variations as SHN_4S_5U_1B^{+++} in Table 9. The architectural variations for 1 band spectrogram (SHN_4S_5U_1B^{+++}) are the same as SH_4S_5U_1B^{+++} which we already did in the ablation experiment of the number of upsampling and downsampling steps. PSHN using 2 bands spectrogram (PSHN_4S_5U_2B^{++}) still achieves the higher result in compare with SHN using 1 band (SHN_4S_5U_1B^{+++}) and PSHN using 4 bands (PSHN_4S_5U_4B). This is because of the huge architectural differences between these bands. For, PSHN_4S_5U_4B, the full band spectrogram of size $F \times T \times 1$ is divided into four equal bands of size ${}^{F}/_{4} \times T \times 1$. The number of parameters and the size of the network increases heavily as five total bands need to pass into the PSHN network separately. This indicates that four band PSHN networks (PSHN_4S_5U_4B) overfit when the network gets deeper despite a small amount of training data for the DSD100 dataset. The experiment is not carried out for 3 band spectrograms as it is not possible to split the spectrogram into three equal halves.

4) THE NUMBER OF STACKS

The experimental results in this study are based on the number of hourglass modules that are stacked together. The SDR effect is measured with no-stack (PHN_1S_4U_2B), 4 stacks (PSHN_4S_5U_2B^{++}), 5 stacks (PSHN_5S_5U_2B), and

6 stacks PSHN (PSHN_6S_5U_2B). All the experimental results are measured using 2 band spectrograms with five downsampling and upsampling steps except the PHN_1S_4U_2B. It can be shown that PSHN with 4 stacks provides a better result compare with 5 stacks, 6 stacks, and no-stack. This, again indicates that PSHN with 5 stacks and 6 stacks overfit when the network gets deeper despite a small amount of training data for the DSD100 dataset. Similarly, it can be inferred that PSHN with no-stack provides comparable results with the baseline SH_4S_4U_1B_W$^+$ [6] which uses a 4-stack network.

VI. CONCLUSION

This study investigated music source separation through a parallel stacked hourglass network. The network was designed for the ability to effciently model both fine local and global spectrogram structure. The experiment was performed using three datasets: Korean traditional music (*Pansori*) dataset, SiSEC 2016 DSD100 dataset and MIR-1K dataset. We constructed the *Pansori* dataset using online and offline resources synthetically and then used with multiband spectrogram representation. The proposed PSHN architecture received multiband mixed spectrogram representation as input and predict the masks from all previous and next parallel hourglass module. The estimated masks thus multiplied with the input to obtain the predicted spectrogram for each source. The predicted spectrograms transform back into the signal using the inverse of short-time Fourier transform. In this procedure, we isolate the sources of *Pansori*, DSD100 and MIR-1K datasets with high quantitative and qualitative results. The results are compared using the median of Signal-to-distortion ratio (SDR) value, source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR) measured in decibels (dB). The experimental results show that the MIR-1K dataset achieves the best performance result while the singing voice separation task of DSD100 achieves the second-best performance in comparison with the current state-of-the-art method. Similarly, the music source separation task of the DSD100 dataset also achieves comparable results in comparison with the current state-of-the-art method. Our *Pansori* dataset carries similar characteristics to most of the traditional songs of a different culture because most of the traditional songs have less number of musical sources. We hope the proposed dataset will help the new researcher to make a similar source separation task even there is no ground truth for each musical component. Similarly, the lessons learned from multiband spectrogram representation would be useful for future work to increase the accuracy of music source separation task. The proposed system for source separation is applicable in education, Karaoke, and automatic transcription of Korean traditional music.

REFERENCES

[1] K. Kokkinakis and P. C. Loizou, "Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2379–2390, 2008.

[2] E. Gómez, F. J. C. Quesada, J. Salamon, J. Bonada, P. Vera, and P. Cabanas, "Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing," in *Proc. 13th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2012, pp. 601–606.

[3] J. R. Zapata and E. Gomez, "Using voice suppression algorithms to improve beat tracking in the presence of highly predominant vocals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 51–55.

[4] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 21–25.

[5] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 106–110.

[6] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Paris, France, Sep. 2018, pp. 289–296.

[7] W.-H. Heo, H. Kim, and O.-W. Kwon, "Source separation using dilated time-frequency DenseNet for music identification in broadcast contents," *Appl. Sci.*, vol. 10, no. 5, p. 1727, Mar. 2020.

[8] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 483–499.

[9] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, Jun. 2000.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2001, pp. 556–562.

[11] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Trans. Neural Netw.*, vol. 16, no. 4, pp. 992–996, Jul. 2005.

[12] M. G. Lopez, H. M. Lozano, L. P. Sanchez, and L. N. O. Moreno, "Blind source separation of audio signals using independent component analysis and wavelets," in *Proc. Conielecomp, 21st Int. Conf. Electr. Commun. Comput.*, Feb. 2011, pp. 152–157.

[13] C. P. Dadula and E. P. Dadios, "A genetic algorithm for blind source separation based on independent component analysis," in *Proc. Int. Conf. Humanoid, Nanotechnol., Inf. Technol., Commun. Control, Environ. Manage. (HNICEM)*, Nov. 2014, pp. 1–6.

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

[15] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Montreal, QC, Canada, Nov. 2017, pp. 1265–1269.

[16] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2135–2139.

[17] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process., Inst. Electr. Electron. Engineers*, vol. 24, no. 10, pp. 1652–1664, Sep. 2016.

[18] P. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. Ismir*, 2014, pp. 477–482.

[19] P.-S. 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.

[20] G.-X. Wang, C.-C. Hsu, and J.-T. Chien, "Discriminative deep recurrent neural networks for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2544–2548.

[21] S. L. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller, "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation," *CoRR*, vol. abs/1709.00611, pp. 1–6, Sep. 2017.

[22] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 261–265.

[23] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," 2018, *arXiv:1806.03185*. [Online]. Available: http://arxiv.org/abs/1806.03185

[24] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," *CoRR*, vol. abs/1711.00541, pp. 696–700, Apr. 2017.

[25] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," 2018, *arXiv:1803.00702*. [Online]. Available: http://arxiv.org/abs/1803.00702

[26] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*. [Online]. Available: http://arxiv.org/abs/1609.03499

[27] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," 2017, *arXiv:1703.09452*. [Online]. Available: http://arxiv.org/abs/1703.09452

[28] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," *CoRR*, vol. abs/1706.07162, pp. 5069–5073, Apr. 2017.

[29] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," 2018, *arXiv:1802.04208*. [Online]. Available: http://arxiv.org/abs/1802.04208

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4700–4708.

[31] Y. R. Pandeya, D. Kim, and J. Lee, "Domestic cat sound classification using learned features from deep neural nets," *Appl. Sci.*, vol. 8, no. 10, p. 1949, Oct. 2018.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[33] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2274–2284.

[34] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, Austin, TX, USA, 2015, pp. 18–25.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

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

[37] A. Liutkus, F. R. Stöter, Z. Rafi, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. LVA/ICA*, 2017, pp. 66–70.

[38] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, p. 6670.

[39] Y.-H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *Proc. ISMIR*, 2013, pp. 427–432.

[40] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. 18th Int. Soc. Music Inf. Retr.*, Suzhou, China, 2017, pp. 1–8.

[41] N. Ono, Z. Koldovsky, S. Miyabe, and N. Ito, "The 2013 signal separation evaluation campaign," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2013, pp. 1–6.

BHUWAN BHATTARAI received the B.S. degree in computer science and information technology from Patan Multiple Campus (an affiliation of Tribhuvan University), Nepal, in 2015, and the M.S. degree in computer science and engineering from Jeonbuk National University, South Korea, in 2019, where he is currently pursuing the Ph.D. degree with the Artificial Intelligence Laboratory.

His research interests include music information retrieval (MIR), image processing, object detection in images, and music source separation.

YAGYA RAJ PANDEYA was born in Banlek, Dadeldhura, Nepal, in 1988. He received the B.E. and M.E. degrees in computer engineering from Pokhara University, Nepal, in 2010 and 2013, respectively.

He was the Head of the Department of Computer Engineering, Dhangadhi Engineering College (NAST), Dhangadhi, Nepal. He joined the Ministry of Home Affairs, Nepal, from 2015 to 2017. He is currently a Ph.D. Fellow with the Fuzzy Logic and Artificial Intelligence Laboratory, Jeonbuk National University, South Korea. His research interests include audio-video information retrieval, audio event detection and localization, emotion engineering, and animal sound behavior analysis.

JOONWHOAN LEE received the B.S. degree in electronic engineering from the University of Hanyang, South Korea, in 1980, the M.S. degree in electrical and electronics engineering from KAIST, South Korea, in 1982, and the Ph.D. degree in electrical and computer engineering from the University of Missouri, USA, in 1990.

He is currently a Professor with the Department of Computer Engineering, Jeonbuk National University, South Korea. His research interests include image and audio processing, computer vision, emotion engineering, and so on.