Circuits, Systems, and Signal Processing (2023) 42:1083–1104
https://doi.org/10.1007/s00034-022-02166-5

High-Resolution Representation Learning and Recurrent Neural Network for Singing Voice Separation

Bhuwan Bhattarai[1] · Yagya Raj Pandeya[2,4] · You Jie[1] · Arjun Kumar Lamichhane[3] · Joonwhoan Lee[1]

Received: 24 January 2022 / Revised: 23 August 2022 / Accepted: 23 August 2022 /
Published online: 14 September 2022

Abstract

Music source separation has traditionally followed the encoder-decoder paradigm (e.g., hourglass, U-Net, DeconvNet, SegNet) to isolate individual music components from mixtures. Such networks, however, result in a loss of location-sensitivity, as low-resolution representation drops the useful harmonic patterns over the temporal dimension. We overcame this problem by performing singing voice separation using a high-resolution representation learning (HRNet) system coupled with a long short-term memory (LSTM) module to retain high-resolution feature map and capture the temporal behavior of the acoustic signal. We called this joint combination of HRNet and LSTM as HR-LSTM. The predicted spectrograms produced by this system are close to ground truth and successfully separate music sources, achieving results superior to those realized by past methods. The proposed network was tested using four datasets (DSD100, MIR-1K, Korean *Pansori*, and Nepal Idol singing voice).

✉ Joonwhoan Lee
chlee@jbnu.ac.kr

Bhuwan Bhattarai
bhubon240@gmail.com

Yagya Raj Pandeya
yagyapandeya@gmail.com

You Jie
youjie80@gmail.com

Arjun Kumar Lamichhane
arjun.lamichhane000@gmail.com

1 Department of Computer Science and Engineering, Jeonbuk National University, Jeonju, South Korea

2 Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal

3 Central Department of Computer Science and Information Technology, Tribhuvan University, Kirtipur, Nepal

4 Deep AI Nepal under Guru Technology, Kathmandu, Nepal

Our experiments confirmed that the proposed HR-LSTM outperforms state-of-the-art networks at singing voice separation when the DSD100 dataset is used, performs comparably to alternative methods when the MIR-1K dataset is used, and separates the voice and accompaniment components well when the *Pansori* and NISVS datasets are used. In addition to proposing and validating our network, we also developed and shared our Nepal Idol dataset.

Keywords Singing voice separation · HRNet · LSTM · Deep neural networks

1 Introduction

Music is a blend of vocal and instrumental sounds to express and evoke emotion through a combination of melody, rhythm, and harmony. The ultimate goal of singing voice separation systems is to separate previously mixed vocal and instrumental components and achieve a deep understanding of each component. These systems, once developed adequately, will have applications in bilateral cochlear implants [15], the ability to calculate fundamental frequency [7], beat monitoring (despite dominant voices) [49], and karaoke music production, as well as any other system that relies on lyric, instrument and chord recognition. Other potential applications include melody extraction/annotation [5, 34], assessment of singing ability [18], automatic lyrics recognition/matching [23, 44], singing visualization [19], and singer identification [20].

The separation of singing voices has long been acknowledged as a difficult task. In recent years, researchers have focused on data-driven machine learning approaches to separate voices from polyphonic music. These systems have generally relied on the two-dimensional time–frequency magnitude spectrogram of the audio signal, which help convolutional neural networks to implement audio-related tasks.

A high-level outline of our proposed singing voice source separation process is presented in Fig. 1. We proposed a long short-term memory (LSTM)-based high-resolution representation network to extract a final feature map from the input spectrogram. This feature map can be used to generate a time–frequency soft mask, which is multiplied with input spectrograms to generate predicted spectrograms. The

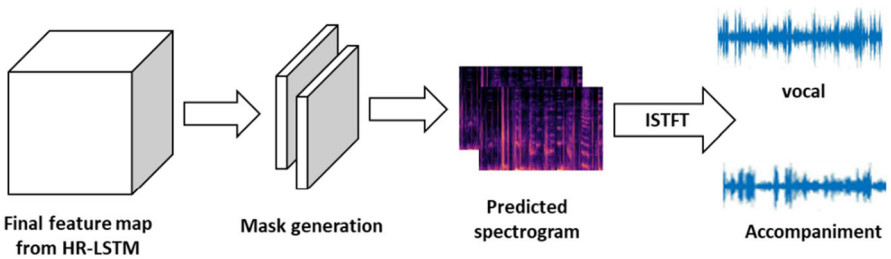

Fig. 1 A basic overview of our singing voice source separation method

predicted spectrograms can be transformed back into signals that correspond to vocal and accompaniment tracks using the inverse of the short-time Fourier transform.

The proposed HR-LSTM network can be employed not only for singing voice separation but also for many different fields of sequence-to-sequence learning problems. The sequence-to-sequence learning includes many different fields of computer vision such as speech recognition, time series prediction, machine translation, and question-answering. The deep neural networks can only be applied to problems whose inputs and outputs are encoded with vectors of fixed dimensionality. However, this LSTM-based HRNet can process not only single data points such as images but also entire sequences of data such as speech and video.

Resolution is important for audio analysis using time–frequency representation because the pixel correlation and harmonic representation in spectrogram are responsible for the unique characteristics of an acoustic signal. We used HRnet; a proven method; to keep the spatial resolution of various feature maps of the network. Temporal information is captured using the LSTM network because the music information is globally correlated along the temporal axis. The consideration of spatiotemporal information without losing the resolution is the main characteristic of our proposed method that is responsible for a better result for music source separation.

The proposed network was tested using three publicly available datasets (DSD100, MIK-1K and *Pansori*). The DSD100 and MIR-1K datasets were used to test how the system separated two music sources, singing voices and accompaniments, while the Korean traditional music *Pansori* dataset was used to test how the system separated two different singing voices, as well as a drum sound. To confirm the utility of the HR-LSTM, we also proposed a new singing voice separation dataset (referred to as NISVS). We mixed the DSD100 and NISVS datasets and reported our results accordingly.

The proposed network's performance was evaluated using the median value of the signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR), each measured in decibels (dB). The proposed HR-LSTM outperformed the current state-of-the-art result, tested against the combined DSD100 and NISVS dataset. The combination of these two datasets in the training phase can be able to predict spectrograms closer to ground truth while testing. The addition of the *Pansori* dataset, along with the incorporation of the NISVS dataset in the DSD100, allowed our model to achieve better separation results than when it was tested only against the DSD100 and MIR-1K datasets.

The major contributions of our study can be summarized as follows:

1. The new model HR-LSTM was proposed and was consisting of a combination of the relatively new "HRnet" for high-resolution representation learning (devised for application to image processing problems) with a well-known approach that uses a long short-term memory (LSTM) module to capture the temporal features of the acoustic signal.
2. We tested our proposed model against various state-of-the-art methods and achieved improvement over the state-of-the-art in certain cases.
3. The new synthetic NISVS dataset was proposed and mixed with the real DSD100 dataset for training. This mixing of two datasets during the training phase improves the result while testing in comparison with without mixing them. This type of

experiment concludes that mixing synthetic and real data while training can improve the accuracy in the test phase for real data in the source separation domain. This can be shown in Tables 3 and 5.

4. HRNet has not been studied yet in the source separation community. So, one of the primary reasons behind the performance of the network is the HRNet itself, which maintains the high-resolution feature map of the spectrogram instead of recovering it from low-resolution, which we explain in the third last paragraph of the "related work" section and the last sentence of "High-resolution representation learning section. This is the main contribution of our paper.

2 Related Work

The study of singing voice separation has a long history. The oldest algorithms written to achieve this goal emphasized the pitch and frequency of the audio signal using statistical methods to separate mixed sources. Independent component analysis (ICA) [13], nonnegative matrix factorization (NMF) [16], and sparse component analysis (SCA) [6] were developed for use with blind source separation [3, 21]. Each of these methods was predicated on the idea that data can be projected from a time series onto a new set of axes based on a statistical technique. The authors of [32] first separated the singing voice from music accompaniment using nonnegative matrix partial co-factorization (NMPCF). The separated singing voice was subsequently used to estimate the pitches and reconstruct the singing voice's spectrum. The authors of [31] attempted to separate music/voice by first identifying the periodically repeating segment from a mixture, and then separating the repeated signal from the mix.

Today, these methods have been supplanted by deep neural networks capable of outperforming previous approaches to source separation. To the latent information from an audio input, deep neural networks often use a hierarchical architecture and a nonlinear approximation function to estimate the independent music source from the combined signals. The authors of [42, 46, 50] applied this deep learning-based separation method. For single-channel source separation, a fully convolutional denoising auto-encoder (CDAE) was presented by [8]. In that case, the researchers explored whether a CDAE could derive the spectral-temporal filters and properties associated with a source. The authors of [24] and [25] similarly used multichannel audio input to train a DNN by focusing primarily on the spectral properties of a single frame. To separate the source spectra, these authors employed a fully linked network and a 2D Mel-Spectrogram. The authors of [35] suggested a modified group delay (MOD-GD) function intended to improve the performance of the available algorithms by incorporating previously neglected phase spectrogram information. DNN was used by the authors of [2] for supervised speech signal training that enhanced speech intelligibility in noisy environments.

Some researchers have used waveform music representation for source separation and to maintain the audio signal's phase information. Preserving the sinusoid audio information [4] and managing the memory requires a data-driven strategy when applied to waveform audio representation. The U-Net-based architecture employed in [38]

resampled the characteristics at various time scales. Through the incorporation of source additivity into the output layer, upsampling, and context-aware prediction, a modified U-Net design outperformed its peers. Similarly, the authors of [9, 22] used an encoder-decoder approach to solve the problem of many speakers requiring multiple audio channel source separation. The authors of [27, 30, 33] also used waveform representation to isolate speech from noisy signals, while in the authors of [1, 11, 39, 40] instead of using waveform audio representation, divided the spectrogram into multiple sub-bands based on frequency ranges to generate the time–frequency masks. As the patterns of the spectrogram were different along with the frequency band, applying a different convolutional filter in each band proved to be critical to boosting the performance of the source separation systems.

None of the aforementioned studies on music information retrieval preserves the high-resolution representation of the spectrogram. HRNet has recently proven successful at human pose estimation, object detection, and semantic segmentation [45]. Accordingly, HRNet has supplanted the encoder-decoder-based networks designed to recover high-resolution from low-resolution. HRNet, which maintains high-resolution instead of recovering them from low to high-resolution results in highly precise and semantically strong features. This preserves the correlation between different Mel-bins of the spectrogram in succeeding layers of HRNet. In this paper, we sought to determine whether maintaining spectrogram features, rather than recovering them in subsequent layers (as is done by other deep neural networks) results in enhanced performance in separating singing voices. To further improve performance at this task, we blended the HRNet with an LSTM block. While the blending of architectures often increases the complexity of the model in an undesirable manner, we combined the two in a unified architecture similar to the authors of [39].

Existing singing voice separation datasets are limited in terms of both size and musical variety. To advance our study and the field more broadly, we created a labeled dataset for the singing voice separation problem based on data from a Nepali reality television singing competition. Despite the fact that the proposed dataset is synthetic, it effectively preserved the real-world music samples, ensuring that HR-LSTM performance was unaffected in real-world test samples.

To further improve the system's performance, the proposed samples of the training dataset were mixed with training samples of DSD100, and our test results are reported independently. As deep learning architectures are only effective when provided with adequate and appropriate data, we compared our methods against prior state-of-the-art alternatives using various publicly available datasets.

2.1 The New Nepal Idol Singing Voice Separation Dataset

Idols is a franchise reality television singing competition created by British television producer Simon Fuller and developed by Fremantle [47]. Nepal Idol is a Nepali reality television singing competition that is part of the Idols franchise. In the Nepal Idol competition, each contestant must pass four different selection rounds (audition, theatre, piano, and gala) prior to proceeding to the final round. Our Nepal Idol singing voice separation dataset (NISVS) was generated using recordings from the

Table 1 The NISVS dataset in detail

Data	Singing voice	Accompaniment	Mixture	Average duration (s)
Training data	70	70	70	32.79
Test data	25	25	25	34.15

audition round. The contestants in this round must perform without any instrumental accompaniment. This allow us to establish ground truth sources for each singer's voice. Likewise, to obtain the ground truth sources for the instruments, we downloaded Nepali instrumental sounds similar to those used in later rounds from YouTube. The downloaded instrumental sounds and the contestant voices without instruments were mixed together in equal length to obtain mixed signals. This process of constructing the synthetic mixtures does not have proper alignment between instruments and singing voices and is uncorrelated with each other. Although the synthetic data is uncorrelated, it can be used to add real data to increase the number of samples just during the training phase. The synthetic data have already been used and perform well in deep learning communities by performing various augmentation technique in image and audio domains. Likewise, the construction of our synthetic NISVS dataset and added it with real training dataset support to increase the accuracy of the real dataset during the testing phase. More specifically, we merge well-aligned real training data of DSD100 with uncorrelated training data of the NISVS dataset. The result of the experiment during the testing phase for both real (DSD100) and synthetic (NISVS) samples are reported in the experimental section of Tables 3 and 5. This result states that the mixing of real and synthetic data during the training phase can improve the accuracy in the test phase in comparison with without mixing it.

In total, we ended up identifying 95 sources, including singing voice, accompaniment and mixture. Seventy of these were placed in the training set while the remaining 25 were placed in a test set. The dataset was preprocessed to ensure that the singing voice and accompaniment recordings were precisely the same length. The audio samples were of varying length, ranging from 10 to 78 s, with an average duration of 32.79 s in the training set. In the test set, samples were 13–75 s with an average duration of 34.15 s. The training data included 2296 s of recordings total, while the test data included 854 s. Consistent with the format of the DSD100 dataset, the two sources and mixture recordings that comprised our NISVS dataset were kept in different folders, [26]. The details of the NISVS dataset are presented in Table 1. The log spectrogram visualization of the two ground truth sources along with their mixture is shown in Fig. 2.

2.2 High-Resolution Representation Learning

The most well-known encoder-decoder based networks (e.g., hourglass, U-Net, DeconvNet, SegNet) were designed to recover high-resolution from low-resolution representation by upsampling the feature map. The distinguishing feature of these

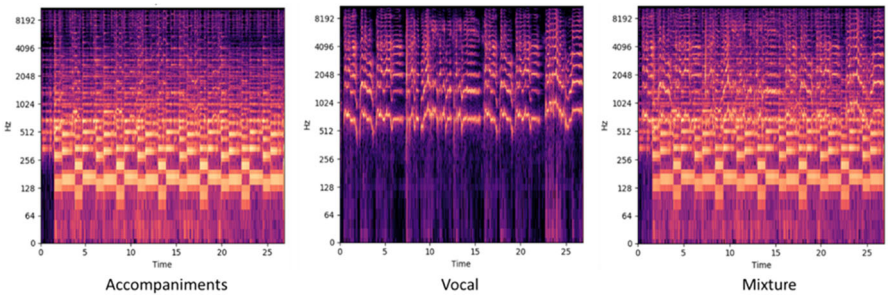

Fig. 2 Log spectrogram visualization of the NISVS **a** accompaniments, **b** vocal, and **c** mixture

networks is that they connect multi-resolution convolutions in series, with the result that the representations are weak due to location-sensitivity loss. HRNet solved this problem by connecting multi-resolution convolutions in parallel. HRNet was originally developed as a backbone network and is currently among the best performing networks at human pose recognition, object detection, and semantic segmentation [45]. The features obtained from HRNet are highly precise and semantically strong as the network maintains the high-resolution instead of recovering it from low to high-resolution. HRNet architecture connects high-to-low-resolution convolution in parallel with repeated fusion instead of series.

The HRNet architecture is made up of four blocks, each of which symbolizes a multi-resolution that connects high-to-low and low-to-high in parallel. Starting with high-resolution in the first block, network processing gradually adds high-to-low resolution one by one. High-to-low resolution is gradually added to create new stages and link the parallel multi-resolution streams. This consists of resolutions from the previous stages and one extra low-resolutions from the current stage. Multi-resolution features are created by fusing these resolutions together. Figure 44 depicts the multi-resolution fusion layer during low-to-high and high-to-low processes. To fully follow its processing, readers are encouraged to view the original paper of HRNet [45].

2.3 LSTM Blocks

Recurrent neural networks (RNNs) are powerful tools used for sequence learning in a diverse array of fields, from speech recognition to image captioning. It can also be used in control theory for non-fragile $\mathcal{H}_\infty$ synchronization. Bidirectional associative memory inertia neural networks has recently been proposed for the synchronization of discrete-time by combining continuous-time inertial neural networks and conventional first-order bidirectional associative memory neural networks [36]. Similarly, the method for synchronization controller has been proposed to handle the controller gain fluctuations in [37]. So, in order to use the power of RNNs, we propose a LSTM block that receives an N size feature map as input and provides $N + 1$ size feature maps as output. The LSTM block used was made up of a 1×1 convolution that reduced the number of feature maps to one. The two-dimensional feature vectors from the single feature map are converted into a one-dimensional feature vector by using the

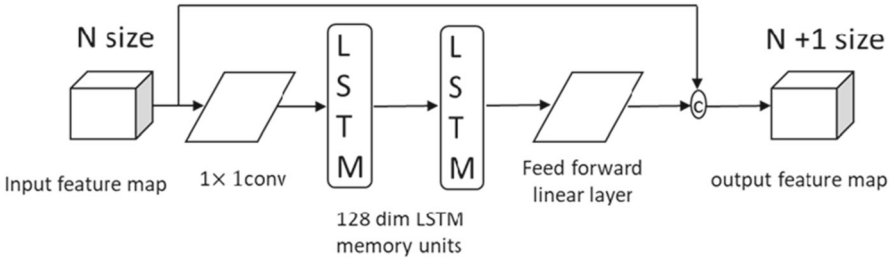

Fig. 3 The LSTM block architecture

LSTM layer. There are two LSTM layers each containing 128 dimension memory units. These memory units in the LSTM block represent the one-dimensional feature vector of the spectrogram. The final layer of the LSTM block is a feedforward linear layer that converted the number of LSTM units back into the input frequency.

This type of recurrent structure has the advantage of capturing the context information from nearby frames of the mixed spectrogram. The context information is important to capture the temporal structure of the mixed-signal from which the network can memorize the longer dependencies and thus helps to improve the singing voice separation results. The LSTM blocks has been adopted in every stage of HRNet (as depicted in Fig. 55) to obtain LSTM-based multi-scale feature map. The features obtained from the HRNet are effective for modeling the local structure of the spectrogram as it follows the CNN structure. Whereas, the features from the LSTM block captures the global information by covering the entire frequency at once. The concatenation of these two multi-scale local and global features is well suited to separate the singing voices. The architecture of the LSTM block is described in Fig. 3.

2.4 Combining LSTM with HRNet

Prior studies on source separation [39, 41] have suggested that the blending of two networks, particularly convolutional neural networks and LSTMs, can increase audio source separation accuracy. These blended networks achieve state-of-the-art results when tested against various publicly available datasets. Inspired by this technique, in this paper we aimed to adapt an HRNet for use with an LSTM to perform singing voice separation.

The input to our HR-LSTM was the mixed magnitude spectrogram of size $F \times T \times 1$, in which $F = 512$ denoted the frequency axis, $T = 64$ denoted the time axis, and 1 was the spectral channel. The HR-LSTM consisted of four branches (branch 1–4) that calculated a high-resolution spectrogram from a sub-network of branch1 in parallel with a lower-resolution spectrogram from sub-networks of branch2, branch3 and branch4. Similar to the ResNet-50, each branch consisted of four residual units with skip connection [10]. The output feature map of the last residual block in all branches was passed into the LSTM block of 128-dimension memory units. The spectrogram feature map obtained from the LSTM block and the residual block were concatenated, resulting in the output as an LSTM-based multi-scale feature map in

all four branches as illustrated in Fig. 55. These concatenated LSTM-based multi-scale feature maps capture the local and global features which are useful to effectively separate the singing voices. After obtaining the feature map, there is a fusion layer, whose objective is to fused downsampled and upsampled features by aggregating the information obtained from high, medium, and low-resolution feature maps.

Figure 4 demonstrates the multi-resolution fusion layer that has been fused to share the information across the different resolution. The feature maps from low resolution to high resolution simply increase the resolution by using bilinear upsampling, whereas the feature maps from high resolution to low-resolution decrease the resolution by using convolutional with a stride of 2. The final feature maps of each branch are obtained by summing all the downsampled and upsampled features, resulting in the high-resolution representation of the mixed spectrogram. This final high-resolution representation of the spectrogram feature map provides a better trade-off between time and frequency resolutions.

We added the LSTM block to the HRNet at a point just prior to the downsampling and upsampling being performed, allowing the block to capture the global structure and the HRNet to model the fine local structure of the input mixed spectrogram. The architecture of HR-LSTM is shown in Fig. 5.

HR-LSTM's first branch, given input mixed spectrograms of size $F \times T \times 1$, produced outputs with a resolution of $F \times T \times (C + 1)$. In this case, $C = 32$, i.e., the number of channels obtained from HRNet, while 1 represented the feature channel obtained from the LSTM. Similarly, the HR-LSTM's second, third and fourth branches

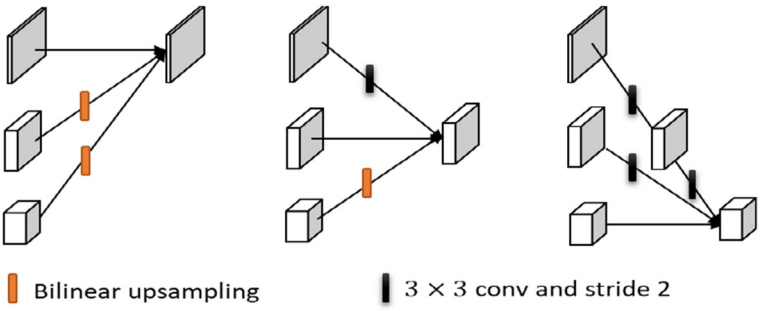

Fig. 4 Multi-resolution fusion layer during low-to-high and high-to-low process

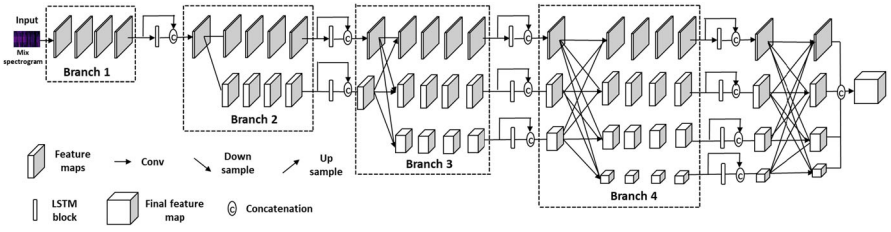

Fig. 5 The architecture of HR-LSTM

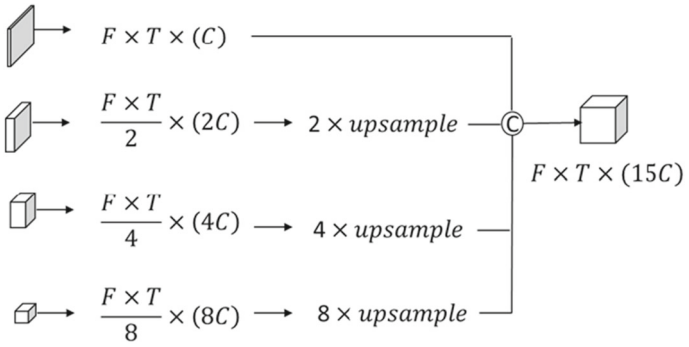

Fig. 6 The multi-resolution outputs were upsampled by a factor of 2, 4, and 8 to obtain the final feature map

gave the output size spectrogram of resolution $\frac{F\times T}{2} \times (2C+1)$, $\frac{F\times T}{4} \times (4C+1)$, and $\frac{F\times T}{8} \times (8C+1)$, respectively. All HR-LSTM branches maintained these resolutions throughout the process. The multi-resolution feature map obtained from the low-to-high and high-to-low processes are fused via the fusion layer to obtain the number of feature maps (C, $2C$, $4C$, and $8C$, for each layer, respectively). The feature of channel $2C$, $4C$, and $8C$ was bilinearly upsampled by corresponding factors of 2, 4, and 8 to obtain feature maps of the same size. Finally, as shown in Fig. 6, all features were concatenated to obtain the final feature map of size $F \times T \times (C + 2C + 4C + 8C)$.

2.5 Loss Function

The error rate for $L2$ loss will be higher because of the squared differences between the predicted and ground truth spectrogram. So, in this work, $L_{1,1}$ norm was used to minimize the absolute difference between the target and predicted spectrograms as it is resistant to outliers in the data which is helpful to effectively ignore the outlier of the spectrogram and has been already studied in various types of source separation problem [1, 29, 40]. The HR-LSTM's time–frequency output mask for the i^{tn} source spectrogram is represented by M_i. The input mixed spectrogram X of size $F \times T \times 1$ was multiplied with M_i to obtain the predicted spectrogram. The loss function for the i^{tn} source spectrogram was used to minimize the absolute difference between the ith ground truth spectrogram of the music source and was given by

$$\text{Loss}_{i\text{th}} = \|Y_i - X \odot M_i\|_{1,1} \tag{1}$$

where $\odot$ is defined as the element multiplication, $\|\cdot\|_{1,1}$ is the 1-norm, and M_i represents the time–frequency mask for the i^{tn} music sources. The prediction of the HR-LSTM prior to the application of the time–frequency mask to the singing voice is represented by $P_{1\text{st}}$ and prior to the application of the time–frequency mask to the accompaniment is represented by $P_{2\text{nd}}$. The time–frequency mask M is defined as

$$\text{Mask}_M = \frac{|P_{1\text{st}}|}{|P_{1\text{st}}| + |P_{2\text{nd}}|}, \tag{2}$$

The output of the network for singing voice and accompaniment is given by

$$\text{singing voice}(P_{1\text{st}}) = \text{Mask}_M \odot X, \tag{3}$$

$$\text{Accompaniment}(P_{2\text{nd}}) = (1 - \text{Mask}_M) \odot X, \tag{4}$$

Equation (1) represents loss for a single source spectrogram. Accordingly, the total loss of the network for N music sources is defined as

$$\text{Loss}_N = \sum_{1=1}^{N} \text{Loss}_{i\text{th}} \tag{5}$$

2.6 Evaluation Measures

The HR-LSTM network was evaluated using the popular metrics of the median of signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR), each measured in decibels (dB) consistent with BSS-Eval metrics [43]. Initially, according to BSS-Eval, it was assumed that the estimation of the predicted sources $\hat{Z}_{predicted}$ was composed of four independent components, as given in Eq. (6).

$$\hat{Z}_{\text{predicted}} = Z_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}, \tag{6}$$

where $\hat{Z}_{\text{predicted}}$ is any source predicted by the network, Z_{target} is the ground truth source, and e_{interf}, e_{noise}, and e_{artif} are error terms for interference, noise, and artifacts [43]. The calculation of all evaluation measures requires knowledge of the ground truth signals divided into short window segment of few seconds long. SIR reflects the number of additional sources that can be heard in the estimated source, whereas SAR describes the number of unwanted artifacts between the true source and predicted source. SDR is used as a general indicator to measure the effectiveness of the source separation system. Equations (7), (8) and (9) measure the SDR, SIR and SAR ratio between the predicted and ground truth signal.

$$\text{SDR} = 10\log_{10}\frac{||Z_{\text{target}}||^2}{||e_{\text{artif}} + e_{\text{interf}} + e_{\text{noise}}||^2}, \tag{7}$$

$$\text{SIR} = 10\log_{10}\frac{||Z_{\text{target}}||^2}{||e_{\text{interf}}||^2}, \tag{8}$$

$$\text{SAR} = 10\log_{10}\frac{||Z_{\text{target}} + e_{\text{interf}}||^2}{||e_{\text{artif}}||^2}, \tag{9}$$

3 Experiments

The HR-LSTM was tested against four different types of source separation datasets. The three experiments on the DSD100, MIR-1K and NISVS datasets were to test singing voice separation, while the Korean traditional *Pansori* dataset was used to test the system's ability to separate the three sources of drum, drummer voice, and singer voice [1]. DSD100 and MIR-1K are publicly available datasets, while NISVS was our created dataset. Our experimental configurations were the same across all four dataset tests as those of [1], with the exception that we increased the number of iterations to 400,000. We also perform experiment by mixing the training data of DSD100 and MIR1K. The mixing of two different datasets achieves slightly better results in comparison with without mixing them.

4 Datasets

As our proposed NISVS dataset was described in Sect. 3, we will briefly review the DSD100, MIR-1K, and *Pansori* dataset. *Pansori* music, which emerged in South Korea, has been registered by UNESCO as an intangible heritage. In this type of music, the singer explains the actions of characters and expresses their feelings during a stage performance. The *Pansori* dataset used in [1] consisted of three different sources to separate; drum, drummer voice, and singer's voice. Drum and drummer voice are repeated throughout the entire song, repeating every 0.5–3 s. The *Pansori* dataset mixed samples were synthetically created by mixing drum, drummer voice and singer's voice with white noise. The sources for drum and drummer voice were initially physically removed from the original *Pansori* song and saved in a different folder, establishing the ground truth source for the singer's voice. The drum and drummer's voice in *Pansori* music only contains the percussive elements, meaning that there are no harmonic and rhythmic elements in *pansori*. So, the synthetic *Pansori* samples used just during the training phase can successfully separate the sources for real *pansori* music in the test phase [1].

The DSD100 dataset, consisting of 100 full-track songs, was originally designed by SiSEC [26]. The dataset consists of an evenly distributed variety of musical genres and styles. Although there are four different sources in the original DSD100 dataset—bass, drum, other, and vocal—the DSD100 dataset used in our experiment was adapted for the singing voice separation task by mixing the bass, drum, and other sources into 'accompaniment.'

The MIR-1K dataset similarly contains 1000 song clips with voice and accompaniment captured in the left and right channels at a sampling rate of 16 kHz. The annotation file of MIR-1K dataset contains additional information, including pitch contours in the semitone, indices and types of unvoiced frames, lyrics, and vocal/non-vocal segment. Each track ranges from 4 to 13 s, and the total length of the dataset is 133 min. The songs were performed by eight women and eleven men, most of whom had no formal music training. To ensure a fair comparison, we selected 175 clips sung by one male (abjones) and one female (amy) as a training set. The remaining 825 tracks were used

for testing our source separation system. Additional details concerning these datasets are shown in Table 2.

4.1 Results of Testing Using the DSD100 Dataset

We merged our proposed NISVS dataset with the training data of DSD100. Then, the median SDR value for HRNet with only DSD100, HR-LSTM with only DSD100 and HR-LSTM with DSD100 plus NISVS has been reported as a test result of DSD100. The DSD100 dataset played two roles: (1) samples with four independent sources had these sources separated, and (2) samples with two independent sources of vocals and accompaniments had singing voices separated. As we designed our experiment specifically for singing voice separation, the three sources of bass, drum, and other were blended. As deep neural networks require large datasets for training, mixing DSD100 and our NISVS dataset slightly improved median SDR values in the HR-LSTM trained on the modified dataset over that trained on the original dataset. Our HR-LSTM with mixing data can even perform better for separating the vocals which is 0.04 dB more in other state-of-the-art networks, including MMDenseLSTM [39]. However, for separating accompaniments, this is reduced by 0.16 dB. Moreover, our HRNet and HR-LSTM achieved comparable results with other current algorithms when only the DSD100 dataset was used (Table 3).

Other state-of-the-art algorithms, like MMDenseNet [40], MMDenseLSTM [39], and PSHN (4-Stack) [1] use multi-band spectrogram input to predict respective sources of music. These algorithms use parallel network architectures to extract features from each band and concatenate the output feature map from each parallel network to estimate the final sources of music. The MMDenseLSTM, which is an improved version of MMDenseNet, is still among the best at separating the accompaniment which is 0.16 dB more as compared to our HR-LSTM trained on DSD plus NISVS. With respect to vocal separation, however our HR-LSTM outperformed all existing methods. Similarly, the authors of [29] created a fully convolutional hourglass network that used a single-band spectrogram to extract the features using a top-down and bottom-up approach. While the SH-4stack [29] and HR-LSTM both use single-band spectrogram, the prior method was 0.90 dB and 0.43 dB less accurate for vocals and accompaniments. BLEND [41] is similar to our method, as it merged two neural network architectures (feed-forward and recurrent) by combining the output using Wiener filtering, though it performed worse than the proposed method. NUG [24] estimated the source spectra by combining the covariance matrix with a deep neural network, though it achieved an accuracy of only 4.55 dB for vocals and 8.90 dB for accompaniments. DeepNMF [17] is a conventional method for audio source separation that utilizes a nonnegative deep neural architecture and achieved only 2.75 dB accuracy for vocals and 8.90 dB for accompaniments.

4.2 Results of Testing Using the MIR-1K Dataset

HRNet and HR-LSTM were tested without mixing our NISVS dataset with the MIR-1K [12] dataset, as the MIR-1K dataset contains enough audio samples for training

Table 2 Details of the DSD100, MIR-1K, and *Pansori* datasets

Data	Singing voice		Accompaniment		Drum		Drummer voice		Mixture	
	Train	Test	Train	test	Train	Test	Train	Test	Train	test
DSD100	50	50	50	50	–	–	–	–	50	50
MIR-1K	175	825	175	825	–	–	–	–	175	825
Pansori	50	15	–	–	50	15	50	15	50	15

Table 3 Median SDR values in decibel (dB) for singing voice separation on DSD100 dataset

Method	Vocals	Accompaniments
DeepNMF[17]	2.75	8.90
NUG [24]	4.55	10.29
BLEND [41]	5.23	11.70
SH-4stack [29]	5.45	12.14
MMDenseNet [40]	6.00	12.10
MMDenseLSTM [39]	6.31	**12.73**
PSHN (4-Stack) [1]	6.01	12.42
HRNet (only_DSD)	5.90	12.37
HR-LSTM(only_DSD)	6.28	12.51
HR-LSTM(DSD_plus_NISVS)	**6.35**	12.57

Bold values indicate the highest accuracy measure

(825). Moreover, the musical genres in our proposed NISVS dataset are too different from those contained in the MIR-1K dataset.

The performance of MIR-1K dataset has been reported to compare with other state-of-the-art algorithms using GNSDR, GSIR, and GSAR for both singing voice and accompaniment. Global normalized SDR (GNSDR), global SIR (GSIR), and global SAR (GSAR) are calculated as weighted means of NSDR, SIR, and SAR, respectively, which is based on BSS-EVAL metrics [28, 43]. Table 4 presents the experimental results achieved by our combined HR-LSTM and HRNet. The HR-LSTM and HRNet as assessed by GSAR and GNSDR at separating singing voice and accompaniments exceeded all other baseline architectures. Specifically, HR-LSTM performed the best as assessed by GNSDR (for accompaniment separation) and GSAR (for singing voice separation). The HRNet performed best as assessed by GSAR at accompaniment separation, though PSHN (4-Stack) [1] was still best at separating singing voices, as it had a GNSDR value of 10.83 dB and GSIR value of 16.54 dB. PSHN (4-Stack) also separated accompaniment well, as assessed by GSIR, outperforming our HR-LSTM by 0.09 dB.

4.3 Results of Testing Using the NISVS Dataset

Table 5 shows the results of three experiments conducted with the NISVS dataset. First, the mixed audio of the DSD100 and NISVS datasets was used for training. The HR-LSTM network trained on this mixed dataset performed well during testing. It helps, of course, that the music in these two datasets is similar in terms of genre. The connections of the multi-resolution convolutions in parallel in the HRNet has the accuracy of 17.59 dB for vocals and 8.04 dB for accompaniments. This experiment on HRNet only uses our NISVS dataset for training.

A second experiment that used the blended LSTM block and HRNet (HR-LSTM) increased performance by 1.23 dB for vocals and 0.78 dB for accompaniments over an HRNet trained only on our NISVS dataset. Incorporation of our NISVS dataset into the

Table 4 GNSDR, GSIR, and GSAR values in decibel (dB) for singing voice separation using the MIR-1K dataset

Method	GNSDR	GSIR	GSAR
Singing voice			
MLRR [48]	3.85	5.63	10.70
U-Net [14]	7.43	11.79	10.42
SH-1stack [29]	10.29	15.51	12.46
SH-2stack [29]	10.45	15.89	12.49
SH-4stack [29]	10.51	16.01	12.53
PSHN (4-Stack) [1]	**10.83**	**16.54**	12.67
HRNet	10.57	16.29	12.53
HR-LSTM	10.46	15.43	**12.70**
Accompaniments			
MLRR [48]	4.19	7.80	8.22
U-Net [14]	7.45	11.43	10.41
SH-1stack [29]	9.65	13.90	12.27
SH-2stack [29]	9.64	13.69	12.39
SH-4stack [29]	9.88	**14.24**	12.36
PSHN (4-Stack) [1]	9.89	14.01	12.65
HRNet	9.94	14.13	**12.95**
HR-LSTM	**9.97**	14.15	12.71

Bold values indicate the highest accuracy measure

Table 5 Median SDR values in decibel (dB) for singing voice separation using our developed NISVS dataset

Method	Vocals	Accompaniments
HRNet (only_ NISVS)	17.59	8.04
HR-LSTM(only_NISVS)	18.82	8.82
HR-LSTM(DSD_plus_NISVS)	**19.46**	**8.85**

Bold values indicate the highest accuracy measure

system improved the HR-LSTM's accuracy for two reasons; first, the incorporation of the LSTM block before every downsampling and upsampling operation captured the global structure for modeling the fine local features of the input mixed spectrogram, and second, the LSTM treated the feature map of the spectrogram as sequential data along the time axis which captured the long range dependencies present in the mixed spectrogram.

The third experiment tested the HR-LSTM on the blended DSD100 and NISVS dataset. This mixed dataset resulted in the most accurate model performance, reaching 19.46 dB for vocals and 8.85 dB for accompaniments.

We also visualize the predicted and ground truth spectrogram for one of the test audio samples of singing voice and accompaniment of the NISVS dataset. In Table 5, the mixing of the real DSD100 dataset and synthetic NISVS dataset gives better

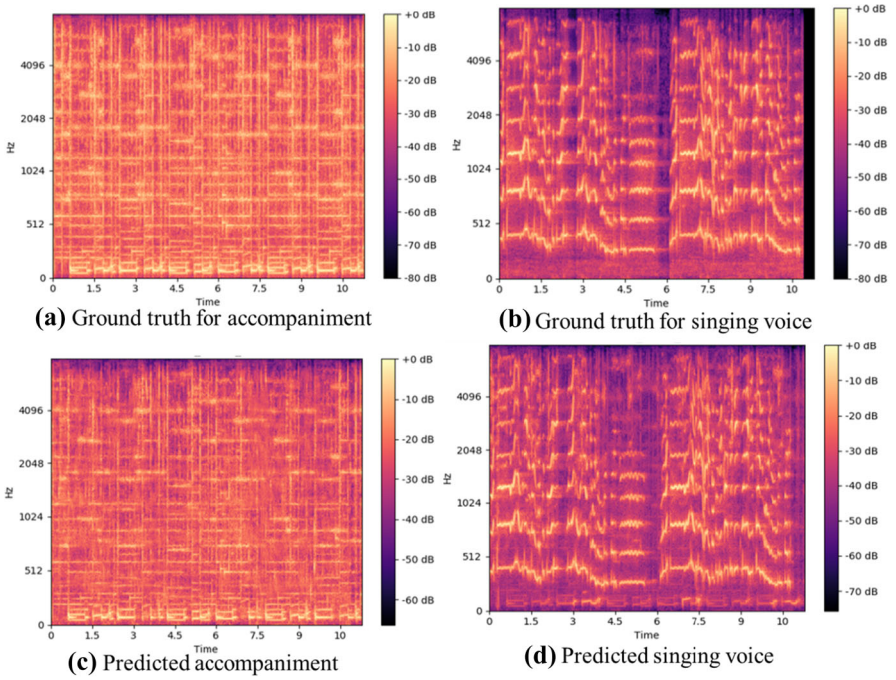

Fig. 7 The results of comparison between ground truth and predicted spectrograms in one of our test set. **a** Ground truth for accompaniment, **b** ground truth for singing voice, **c** predicted accompaniment, **d** predicted singing voice

accuracy in comparison with without mixing it. So, the visualization of the test result in Fig. 7 has been carried out using the HR-LSTM model trained with a mixing DSD100 and NISVS datasets. The visualization results prove that the predicted spectrogram for singing voice and accompaniment are close to ground truth and hence separate the music sources successfully.

4.4 Results of Testing Using the *Pansori* Dataset

Table 6 shows the outcomes of our study, as well as a comparison to the baseline [1] using *Pansori* dataset. The *Pansori* dataset was originally published in [1], in which a parallel stack hourglass network (PSHN) with different architectural variations—PSHN (1-Stack), PSHN (2-Stack), PSHN (3-Stack), and PSHN (4-Stack) – was constructed. The PSHN (4-Stack) performed the best, at 15.97 dB for drum, 12.86 dB for drummer voice, and 16.12 dB for singer voice. The masks that had been estimated in the intermediate stage of the parallel hourglass module were passed into the next module, which resulted in the excellent performance of the PSHN (4-Stack). The HRNet and HR-LSTM architectures designed in this paper surpassed even the PSHN (4-Stack) in accuracy. The HR-LSTM exceeded the HRNet along with all variants of the baseline at drum, drummer voice and singer voice separation. The outperformance

Table 6 Median SDR values in decibel (dB) for the *Pansori* source separation dataset

Method	Drum	Drummer voice	Singer voice
PSHN (1-Stack) [1]	15.65	12.40	15.76
PSHN (2-Stack) [1]	15.81	12.54	15.94
PSHN (3-Stack) [1]	15.89	12.66	16.03
PSHN (4-Stack) [1]	15.97	12.86	16.12
HRNet	15.83	12.91	16.25
HR-LSTM	**16.23**	**13.41**	**16.91**

Bold values indicate the highest accuracy measure

is attributed to the fact that our HRNet connected multi-resolution convolutions in parallel, and the fusion of the HRNet with a LSTM block. Even the HRNet without the LSTM block performed well compared to the baseline, achieving 15.83 dB for drum, 12.91 dB for drummer voice, and 16.25 dB for singer voice (0.14 dB less for drum, 0.05 dB more for drummer voice, and 0.03 dB more for singer voice than the PSHN (4-Stack)).

5 Discussion

The research in [17] makes use of nonnegative deep networks, from which the nonnegative parameters were produced via source separation based on nonnegative factorization. This results from unfolding NMF iterations by untying its parameters. Similarly, the work in [24] combine deep neural networks with spatial covariance matrices to do source separation. In addition to this, [48] propose an algorithm called MLRR to learn the subspaces using online dictionary learning. All these methods [17, 24, 48] uses common approach called matrix factorization which is the conventional and old approach for source separation and is far beyond in compare with our deep learning-based method. Moreover, we compare our proposed HR-LSTM network with other deep learning-based methods [1, 29, 39–41]. The work in [1, 39, 40] uses multi-band spectrogram as input and feed each band input into the separate network. Whereas, in our case, we use single-band spectrogram which makes our architecture less complex in compare with them. In addition to this, the key advantage of our method in compare with all other methods in Tables 3, 4, and 6 is that our method can preserves the high-resolution representation of the spectrogram rather than recovering it from low-resolution. This process of maintaining the spectrogram brings highly precise and semantically strong features. The mixing of our proposed NISVS dataset with publicly available DSD100 dataset while training in our method is another advantage for improving the accuracy in test data in compare with other methods in the literature.

6 Conclusion

We developed an LSTM and HRNet-based high-resolution representation learning method to perform singing voice separation task. HR-LSTM connects multi-resolution convolution in parallel instead of series which helps to maintain the resolution of the spectrogram throughout the whole process. In our unified design the blending of HRNet and LSTM blocks received mixed spectrogram representations as input and predicted the masks for each source. The predicted mask was then multiplied with the input spectrogram to obtain an estimated spectrogram, which was then transformed back into the signal using the inverse of the short-time Fourier transform. The signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR) values were used to determine the system's accuracy in decibels (dB). We validated the HR-LSTM architecture using four datasets: NISVS, DSD100, MIR-1K, and the Korean traditional music (*Pansori*). Our experiments confirmed that the developed HR-LSTM outperforms state-of-the-art networks at singing voice separation when the DSD100 dataset is used, and performs comparably well when the MIR-1K dataset is used. To further boost performance, we combined the DSD100 and NISVS training datasets, and the test results were presented separately. This newly developed NISVS dataset will assist future researchers working on the problem of voice separation, just as our HRNet will, we anticipate, prove useful in applications that require voice separation.

Funding This work was supported by the National Research Foundation of Korea (NRF) under the Development of AI for Analysis and Synthesis of Korean *Pansori* NRF-2021R1A2C2006895 Project. We would like to express our gratitude to the editors of the Writing Center at Jeonbuk National University for their skilled English-language assistance.

Data availability The authors declare that all training and testing data and codes supporting this study are available from the corresponding author upon reasonable request. All other data supporting this study are available within the article. Our proposed NISVS dataset is publicly available in https://github.com/pratikshaya/singing_voice_separation

References

1. B. Bhuwan, R.P. Yagya, L. Joonwhoan, Parallel stacked hourglass network for music source separation. IEEE Access **8**, 206016–206027 (2020). https://doi.org/10.1109/ACCESS.2020.3037773
2. J. Chen, Y. Wang et al., Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. J. Acoust. Soc. Am. **139**(5), 2604–2612 (2016). https://doi.org/10.1121/1.4948445

3. C. P. Dadula, E. P. Dadios, A genetic algorithm for blind source separation based on independent component analysis, in *2014 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, pp. 1–6. IEEE. https://doi.org/10.1109/HNICEM.2014.7016226
4. C. Donahue, J. McAuley, M. Puckette, Adversarial audio synthesis, ICLR 2019. https://doi.org/10.48550/arXiv.1802.04208.
5. Z.C. Fan, J.S.R. Jang, C.L. Lu, Singing voice separation and pitch extraction from monaural polyphonic audio music via DNN and adaptive pitch tracking, in *IEEE International Conference on Multimedia Big Data (2016)*. https://doi.org/10.1109/BigMM.2016.56
6. P. Georgiev, F. Theis, A. Cichocki, Sparse component analysis and blind source separation of underdetermined mixtures. IEEE Trans. Neural Netw. **16**, 992–996 (2005). https://doi.org/10.1109/TNN.2005.849840
7. E. Gómez, F. Canadas, J. Salamon, J. Bonada, P. Vera, P. Cabanas, Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing, in *13th International Society for Music Information Retrieval Conference (ISMIR 2012)*.
8. E.M. Grais, M.D. Plumbley, Single channel audio source separation using convolutional denoising autoencoders, in *Proceedings of the IEEE GlobalSIP Symposium on Sparse Signal Processing and Deep Learning, 5th IEEE Global Conference on Signal and Information Processing (GlobalSIP 2017)*, 14–16 Nov. Montreal, Canada. https://doi.org/10.1109/GlobalSIP.2017.8309164
9. E.M. Grais, D. Ward, M.D. Plumbley, Raw multi-channel audio source separation using multiresolution convolutional auto-encoders, in *26th European Signal Processing Conference (EUSIPCO)*, 2018. https://doi.org/10.23919/EUSIPCO.2018.8553571
10. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, CA, USA, 27–30 June 2016; pp. 770–778. https://doi.org/10.48550/arXiv.1512.03385
11. W.H. Heo, H. Kim, O.W. Kwon, Source separation using dilated time-frequency DenseNet for music identification in broadcast contents. Appl. Sci. (2020). https://doi.org/10.3390/app10051727
12. C.L. Hsu, J.S.R. Jang, On the improvement of singing voice separation for monaural recordings using MIR-1K dataset. IEEE Trans. Audio Speech Lang. Process. (2010). https://doi.org/10.1109/TASL.2009.2026503
13. A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications. Neural Netw. **13**, 411–430 (2000). https://doi.org/10.1016/S0893-6080(00)00026-5
14. A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, T. Weyde, Singing voice separation with deep U-Net convolutional networks, in *18th International Society for Music Information Retrieval Conferencing*, Suzhou, China (2017).
15. K. Kokkinakis, P.C. Loizou, Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients. J. Acoust. Soc. Am. **123**(4), 2379–2390 (2008). https://doi.org/10.1121/1.2839887
16. D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in *Proceedings of the Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, 3–8 December 2001; pp. 556–562.
17. J.L. LeRoux, J.R. Hershey, F.J. Weninger, Deep NMF for speech separation, in *Proceedings of ICASSP*, 2015, p. 6670. https://doi.org/10.1109/ICASSP.2015.7177933
18. K.W.E. Lin, H. Anderson, M.H.M. Hamzeen, S. Lui, Implementation and evaluation of real-time interactive user interface design in self-learning singing pitch training apps, in *Joint Proceedings of International Computer Music Conference (ICMC) and Sound and Music Computing Conference (SMC)* 2014. http://hdl.handle.net/2027/spo.bbp2372.2014.257
19. K.W.E. Lin, H. Anderson, N. Agus, C. So, S. Lui, Visualising singing style under common musical events using pitch-dynamics trajectories and modified traclus clustering, in *International conference on machine learning and applications (ICMLA)*, pp 237–242 (2014). https://doi.org/10.1109/ICMLA.2014.44
20. K. W. E. Lin, T. Feng, N. Agus, C. So, S. Lui, Modelling mutual information between voiceprint and optimal number of mel-frequency cepstral coefficients in voice discrimination, in *International conference on machine learning and applications (ICMLA)*, pp 15–20 (2014). https://doi.org/10.1109/ICMLA.2014.9
21. P.M.G. Lopez, H.M. Lozano, F.L.P. Sanchez, L.N. Oliva, Blind Source Separation of audio signals using independent component analysis and wavelets, in *CONIELECOMP 2011, 21st International*

Conference on Electrical Communications and Computers, pp. 152–157. IEEE. https://doi.org/10.1109/CONIELECOMP.2011.5749353
22. Y. Luo, N. Mesgarani, Tasnet: time-domain audio separation network for real-time, single-channel speech separation. CoRR (2017). https://doi.org/10.1109/ICASSP.2018.8462116
23. A. Mesaros, T. Virtanen, Automatic recognition of lyrics in singing. EURASIP J. Audio Speech Music Process **1**, 546047 (2010)
24. A.A. Nugraha, A. Liutkus, E. Vincent, Multichannel music separation with deep neural networks, in *Proceedings of EUSIPCO* (2015). https://doi.org/10.1109/EUSIPCO.2016.7760548
25. A.A. Nugraha, A. Liutkus, E. Vincent, Multichannel audio source separation with deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. Inst. Electr. Electron. Eng. **24**(10), 1652–1664 (2016). https://doi.org/10.1109/TASLP.2016.2580946
26. N. Ono, Z. Koldovsky, S. Miyabe, N. Ito, The 2013 signal separation evaluation campaign, in *Proc. MLSP*, pp. 1–6 (2013). https://doi.org/10.1109/MLSP.2013.6661988
27. A.V.D. Oord, S. Dieleman, et al., Wavenet. A generative model for raw audio, in *Proceedings of 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 125 (2016).
28. A. Ozerov, P. Philippe, F. Bimbot, 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. **15**(5), 1564–1578 (2007). https://doi.org/10.1109/TASL.2007.899291
29. S. Park, T. Kim, K. Lee, N. Kwak, Music source separation using stacked hourglass networks, in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 23–27 September (2018), pp. 289–296.
30. S. Pascual, A. Bonafonte, J. Serra, SEGAN: Speech enhancement generative adversarial network, in *Conference of the International Speech Communication Association, INTERSPEECH* (2017). https://doi.org/10.48550/arXiv.1703.09452
31. Z. Rafii, B. Pardo, Repeating pattern extraction technique (repet): A simple method for music/voice separation. IEEE Trans. Audio Speech Lang. Process. **21**(1), 73–84 (2012). https://doi.org/10.1109/TASL.2012.2213249
32. B. Raj, P. Smaragdis, M. Shashanka, R. Singh, Separating a foreground singer from background music, in *Proceedings of International symposium on Frontiers of Research in Speech and Music* (2007), pp. 8–9.
33. D. Rethage, J. Pons, X. Serra, A wavenet for speech denoising, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018). https://doi.org/10.1109/ICASSP.2018.8462417
34. J. Salamon, R.M. Bittner, J. Bonada, J.J. Bosch, E. Gómez, J.P. Bello, An analysis/synthesis framework for automatic F0 annotation of multitrack datasets, in *International Society for Music Information Retrieval Conference* (2017).
35. J. Sebastian, H. A. Murthy, Group delay based music source separation using deep recurrent neural networks, in *2016 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, (2016), pp. 1–5. https://doi.org/10.1109/SPCOM.2016.7746672
36. H. Shen, Z. Huang, Z. Wu, J. Cao, J.H. Park, Nonfragile synchronization of BAM inertial neural networks subject to persistent dwell-time switching regularity. IEEE Trans. Cybernet. **52**(7), 1 (2022). https://doi.org/10.1109/TCYB.2021.3119199
37. H. Shen, X. Hu, J. Wang, J. Cao, W. Qian, Non-fragile synchronization for Markov jump singularly perturbed coupled neural networks subject to double-layer switching regulation. IEEE Trans. Neural Netw. Learn. Syst. Early Access (2021). https://doi.org/10.1109/TNNLS.2021.3107607
38. D. Stoller, S. Ewert, S. Dixon, Wave-u-net: a multi-scale neural network for end-to-end audio source separation, in *19th International Society for Music Information Retrieval Conference (ISMIR 2018)*. https://doi.org/10.48550/arXiv.1806.03185
39. N. Takahashi, N. Goswami, Y. Mitsufuji, MMDENSELSTM: an efficient combination of convolutional and recurrent neural networks for audio source separation, in *Proceedings of 16th International Workshop Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan (2018), pp. 106–110. https://doi.org/10.1109/IWAENC.2018.8521383
40. N. Takahashi, Y. Mitsufuji, Multi-scale multi-band DenseNets for audio source separation, in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 15–18 October 2017, pp. 21–25. https://doi.org/10.1109/WASPAA.2017.8169987

41. S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp. N. Takahashi, Y. Mitsufuji, Improving music source separation based on deep neural networks through data augmentation and network blending, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2017), pp. 261–265. https://doi.org/10.1109/ICASSP.2017.7952158
42. S. Uhlich, F. Giron, Y. Mitsufuji, Deep neural network based instrument extraction from music, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE (2015), pp. 2135–2139. https://doi.org/10.1109/ICASSP.2015.7178348
43. E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006). https://doi.org/10.1109/TSA.2005.858005
44. Y. Wang, M. Y. Kan, T. L. Nwe, A. Shenoy, J. Yin, Lyrically: automatic synchronization of acoustic musical signals and textual lyrics, in *ACM International Conference on Multimedia*. ACM, Cambridge, pp 212–219 (2004). https://doi.org/10.1109/TASL.2007.911559
45. J. Wang, K. Sun et al., Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2020). https://doi.org/10.1109/TPAMI.2020.2983686
46. F. Weninger, J. R. Hershey, J. Le. Roux, B. Schuller, Discriminatively trained recurrent neural networks for single-channel speech separation, in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE (2014), pp. 577–581. https://doi.org/10.1109/GlobalSIP.2014.7032183
47. Wikipedia, https://en.wikipedia.org/wiki/Idols_(franchise)
48. Y.H. Yang, Low –Rank representation of both singing voice and music accompaniment via learned dictionaries, in *ISMIR*, pp. 427–432 (2013)
49. J. R. Zapata, E. Gomez, Using voice suppression algorithms to improve beat tracking in the presence of highly predominant vocals, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 51–55. IEEE. https://doi.org/10.1109/ICASSP.2013.6637607
50. H. Zhang, X. Zhang, S. Nie, G. Gao, W. Liu, A pairwise algorithm for pitch estimation and speech separation using deep stacking network, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE (2015), pp. 246–250. https://doi.org/10.1109/ICASSP.2015.7177969