BASS LINE IS THE BASELINE: BASS INSTRUMENT SEPARATION

Hyunsung Cho
KAIST CS
hscho122@kaist.ac.kr

Youngbo Shim
KAIST CS
aram6207@kaist.ac.kr

Wonjong Ryu
KAIST EE
wonjong@kaist.ac.kr

ABSTRACT

In band music, bass plays an important role in chord progression and melody construction of other instruments. In order to follow the performance of a band, the sound of the bass must be dictated for music and chord transcription. However, since the bass sound often lie in the low frequency band and overlap with the drumbeats, it is hard to get an exact note of it by hearing. In this project, bass instrument separation is implemented based on CNN-based neural networks. There are three contributions developed by our team: 1) multi-channel inputs including the original mixture source and low pass filter (LPF) fed sources, 2) a customized loss function for a bass source separation, 3) a data augmentation to improve the performance. As a result, the multi-channel architecture using LPF approach appeared to be most effective (SDR: -1.54db). Also, through a qualitative assessments we found that the high frequency instrumental sound was well removed, especially the vocal sound.

1. INTRODUCTION

Music source separation has drawn the attention of many researchers in the past few years, with noticable performacnce improvements by the means of Deep Learning(DL). While being an interesting problem in itself, the separation of sources from a mixture can serve as a intermediary step for other tasks such as automatic speech recognition and fundamental frequency estimation.

As mentioned above, DL is a paradigm that shows innovative performance not only in the field of music but also in all fields such as image processing, natural language problem. There are a number of studies that have introduced DL into music source separation. The state-of-the-art model is MM-denseNet [5], which is consisted of multi-scale and multi-band blocks.

In this project, DL was used to implement bass instument separation, especially CHA model [1] was used and modified to take multi-channel inputs. MUSDB18 [3] was used as a dataset to train and test the network, and four data augmentation techniques were given to make further variations to the dataset. Then the augmented data is passed through two types of LPF. Finally, the original source and two different LPF-fed source are used as input. The result is evaluated with a given museval [4] tool.

In Section 2, previous studies on music source separation are introduced. The methodology of this project including the data augmentation, LPF, and customized loss function are proposed in Section 3. Finally, the result is shown in Section 4 and limitation and future work is discussed in Section 5.

2. RELATED WORK

Several techniques have been proposed for source separation of musical audio. Since deep learning techniques were introduced in the field, studies have been competing for performance by modifying the structure of the network.

MM-DenseNet [5] combined two types of network which are multi-scale band and multi-frequency band. The multi-scale band enables the network to model both long contexts and fine-grained structures efficiently. The multi-band train each frequency band separately so that information of high frequency components is not ignored.

Stacked-Hourglass model [2] captures both holistic features from low resolution feature maps and fine details from high resolution feature maps. As passing through multiple modules, the results are refined and intermediate supervision helps faster learning in the initial state.

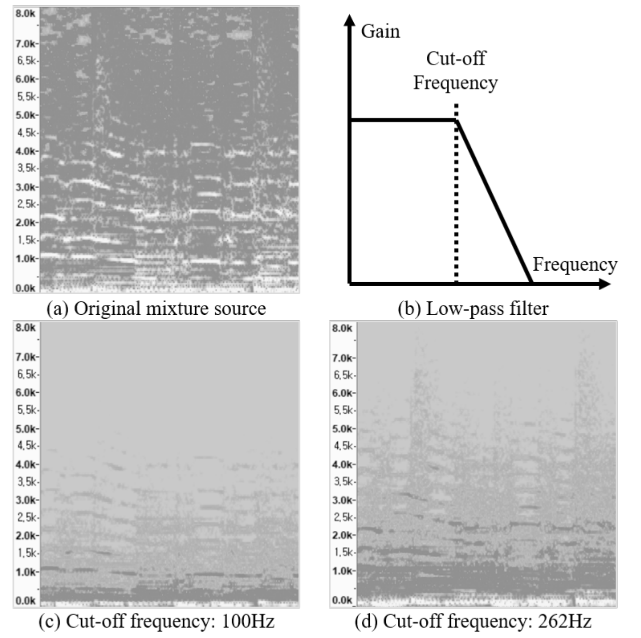
Blend model [6] proposed a blended architecture of Feed-forward Neural Networks (FNN) and Bidirectional Long-Short Term Memory (BLSTM) systems. Raw outputs of both models are combined and a multi-channel Wiener filter post-processing is applied to the final spectrogram.

CHA model [1] divided the role of each layer into vertical convolution layer and horizontal convolution layer. Vertical convolution layer captures local timbre information while the horizontal convolution layer models temporal evolution for four instrumental sources. This model was used in the project due to its flexibility for modifying the architecture and the loss function.

3. METHODOLOGY

3.1 Dataset

MUSDB18 [3] dataset contains 150 music tracks composed of 5 stereo streams (mixtures, vocals, drums, bass, others). 100 music tracks are assigned for training, and 50 music tracks for testing. Furthermore, a given museval [4] is used as an evaluation tool for source separation results.



(a) Original mixture source (b) Low-pass filter

(c) Cut-off frequency: 100Hz (d) Cut-off frequency: 262Hz

Figure 1. (a) Original mixture source (b) Low-pass filter that passes signals with a frequency lower than a selected cut-off frequency (c) spectrogram of the LPF fed signal with cut-off frequency of 100Hz (d) spectrogram of the LPF fed signal with cut-off frequency of 262Hz.

3.2 Data augmentation

Data augmentation was implemented for better performances. Four types of data augmentation techniques are referred from Uhlich et al. [6]'s study: 1) Random swapping left/right channel for each instrument, 2) Random scaling with uniform amplitudes from [0.25 1.25], 3) Random chunking into sequences for each instrument, 4) Random mixing of instruments from different songs. As a result, 500 tracks of length of 30 seconds are obtained.

3.3 Low pass filter

Low-Pass Filter (LPF) is a filter that passes signals with a frequency lower than a selected cut-off frequency and attenuates signals with frequencies higher than the cutoff frequency. In order to implement the bass source extraction effectively, cut-off frequency of 100 Hz and 262 Hz (Figure 1) is selected with two rationales. First, the cut-off frequency of 100 Hz could minimize the vocal sound, which is most dominant in the mixture stream, and collect pure bass sound profile. Cut-off frequency of 262 Hz covers a full bass pitch range (41–262 Hz).

3.4 Network Architecture

A CNN-based neural network was used in this project to develop music source separation as shown in Figure 2. The whole architecture consists of an input stage, an encoding stage, and a decoding stage. The data format of input and output is spectrogram calculated by Short-Term Fourier Transform (STFT).

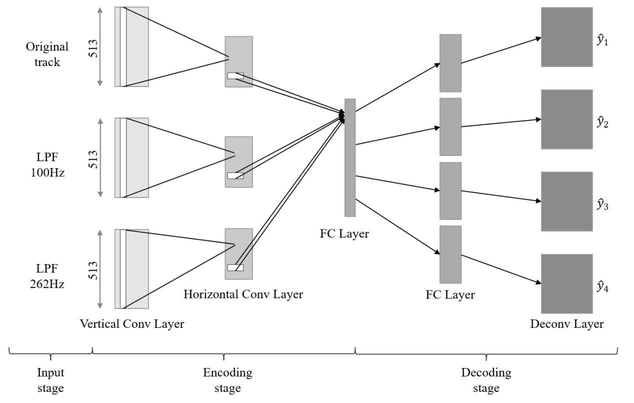


Figure 2. Architecture of the network.

The input of the network are three spectrograms of the mixture source. At this time, the original sound source and the spectrogram of the sound source passed through the LPFs having a cut-off frequency of 100 Hz and 262 Hz are used as inputs.

The input passes through the vertical convolutional layer and the horizontal convolutional layer. The vertical convolutional layer extracts the local timbre information, and the horizontal convolution layer extracts the temporal evolution. After passing through the fully connected layer, it outputs the same size as the input spectrogram through the deconvolution layer. At this time, the spectrograms for vocals, drums, bass, and other instruments are extracted separately.

3.5 Loss Function

The neural network is trained to optimize parameters using a Stochastic Gradient Descent (SGD) with AdaDelta algorithm in order to minimize $L_{sq'}$, the squared error between the estimate source $\tilde{y}_n$ and the original source y_n as described in Eqn (1).

$$L_{sq'} = \sum_{n=1}^{N-1} \parallel \tilde{y}_n - y_n \parallel^2 \tag{1}$$

A few types of loss functions were added such as L_{diff}, representing the difference between the estimated sources, which is described in Eqn (2), and L_{bass}, encouraging differences between bass and other instruments, which is described in Eqn (3).

$$L_{diff} = \sum_{i=1}^{N-1} \parallel \tilde{y}_n - \tilde{y}_{n\neq n} \parallel^2 \tag{2}$$

$$L_{bass} = \sum_{n=1, n\neq bass, drum}^{N-1} \parallel \tilde{y}_n - y_{bass} \parallel^2 \tag{3}$$

Also, we noted that the 'bass' source usually overlaps with the 'drum' source in respect of beat and frequency. To emphasize the difference between these two sources in the separation stage, a $L_{bassdrum}$ loss element, which represents the difference between the estimated bass and drum,

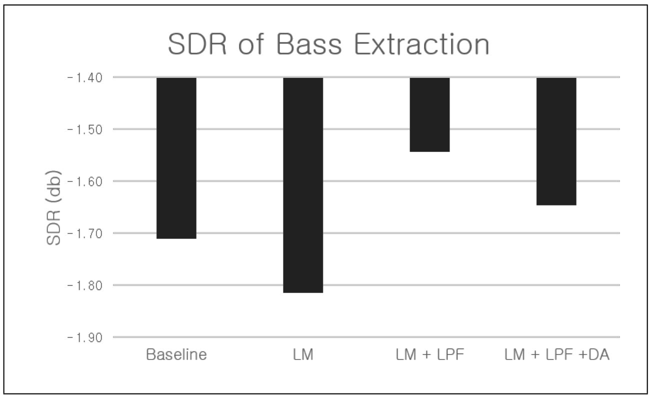


Figure 3. Evaluated SDR of networks trained with various datasets.

was introduced as in Eqn (4).

$$L_{bassdrum} = \| \tilde{y}_{drum} - y_{bass} \|^2 \tag{4}$$

Finally, total loss function is combined as Eqn (5).

$$L_{total} = L_{sq'} - \alpha L_{diff} - \beta L_{bass} - \beta_{drum} L_{bassdrum} \tag{5}$$

4. RESULT AND ANALYSIS

Based on the above methods, four types of network models were developed and then they were evaluated using the given museval [4] tool. A *baseline* model is the basic model without a loss function modification, LPFs, and a data augmentation, which is the same as the CHA model chandna2017monoaural only with a different dataset (MUSDB18 instead of DSD100). The second model (*LM*) is modified with a customized loss function as described in Section 3.5 with only a single input source of the original mixture. *LM+LPF* is a model with a multi-channel input after feeding the LPF to the input source in addition to the *LM* model. The last model (*LM+LPF+DA*) is trained with an augmented dataset (Section 3.2), sustaining the network in *LM+LPF* model. The results of the network models for four metrics (SDR, SIR, SAR, ISR) are shown in Table 1.

Figure 3 shows the changes in SDR, which is an evaluation index similar to human hearing among four metrics. *LM* is a modified version of the loss function for bass extraction, but the result is worse than the base line. The *LM+LPF* is an addition of the LPF model, and the result shows great improvement. LPF would have been affected, but the data was tripled so that it acts like a data augmentation. *LM+LPF+DA* adds data augmentation to the third model, but the result is worse than the third model. Through a random sample listening assessment, we have found that the instruments in high frequency band are effectively suppressed from the mixture source for all models.

model	metric	vocal	others	drum	**bass**
baseline	SDR	-3.41	0.45	1.89	**-1.71**
	SIR	-2.81	0.30	2.40	**-1.77**
	SAR	6.17	1.45	5.81	**5.85**
	ISR	7.17	1.22	8.06	**6.69**
LM	SDR	3.46	0.47	2.01	**-1.81**
	SIR	-2.89	-0.01	2.68	**-1.88**
	SAR	6.22	1.91	5.49	**6.20**
	ISR	6.85	1.26	7.53	**7.15**
LM+LPF	SDR	-3.41	0.43	1.92	**-1.54**
	SIR	-2.58	0.24	2.58	**-1.25**
	SAR	5.84	1.78	5.76	**5.87**
	ISR	6.76	1.26	8.25	**6.81**
LM+LPF+DA	SDR	-4.06	0.32	2.16	**-1.65**
	SIR	-3.32	-0.54	3.55	**-0.95**
	SAR	6.38	1.05	4.80	**5.12**
	ISR	7.14	1.11	8.33	**7.17**

Table 1. Result of Bass Instrument Separation

5. DISCUSSION

5.1 Limitation

The structure of the network was brought from Chadna et al. [1]'s work and modified to fit our input, but the original model itself performs worse than the state-of-the-art models such as MM-DenseNet [5]. The simplest model was selected because it was easy to modify the network and it had a short learning time compared to other complicated models.

The cut-off frequency of applied LPF were 100 Hz and 262 Hz, but it would be better if various LPF were used. However, if too many inputs are given, learning may take too long time or even may not converge. A further study on appropriate frequency band conditions is needed.

Data augmentation was used to increase the number of data from 100 to 500, but in fact the total length of the source became shorter since the augmented track fall short in 30 sec. 2000 data of 30 seconds are prepared, and it could be created as many as we wanted, but only 500 data was used due to the long training time. A better performance is expected when appropriate number of data is used [6].

We have found that instruments in the 'others' source, which are the rest instruments other than 'vocals', 'drums', or 'bass', were still remaining in the predicted bass source. This may have been caused by a deleted loss term from the original model [1], which was used to emphasize the difference between 'others' source with other instruments. We expect that the performance would upturn by summing up the loss terms from the original model and our model.

5.2 Future Work

In this model, the encoder and decoder consist of two types of CNN layers, with a fully-connected layer between them. If the number of layers were increased to 10 or more, the performance could rise with a greater number of data.

Futhermore, another future approach to improve the performance is using U-Nets or residual structures that have been proven to be effective in the vision domain and vocal separation.

Using more various types of filters other than LPF is also a potential future work. Using band pass filter is one option. Since the frequency of the bass starts at 41 Hz, setting the cut-off frequency to 100 Hz means that half of the frequency information is unnecessary. Also, using a longer window for the STFT or employing constant Q-filter bank might be beneficial for extracting the bass line since the low-frequency level details are smudged with standard STFT windows.

In data augmentation, random chunking and swapping were applied to the sound source. This may have resulted in a dataset with independent beat sequences among instruments. Since usual musical audio possess a universal beat and overlapped bass and drum due to this universal beat could have negatively affected bass extraction, a delicate mixing of sound sources by matching the beat may construct a more realistic data.

6. CONCLUSION

In this project, we created a network that extracts bass sound sources from a multi-source audio mixture. We have used three input channels that use the original sound source and signals fed to a LPF with cut-off frequencies of 100 and 262 Hz. We augmented the dataset into 500 sound sources of length of 30 sec using data augmentation techniques. We designed a network's loss function to highlight the bass sound. As a result, the evaluation confirms that using the LPF as input is effective in improving the performance of bass source separation.

7. AUTHOR CONTRIBUTIONS

"Hyunsung Cho" developed the network architecture and evaluations and ran evaluations. "Youngbo Shim" decoded the dataset and coded the data augmentation techniques. "Wonjong Ryu" coded the low pass filter algorithm and drafted the final report. We did previous works search and developed ideas together.

8. REFERENCES

[1] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International conference on latent variable analysis and signal separation*, pages 258–266. Springer, 2017.

[2] Sungheon Park, Taehoon Kim, Kyogu Lee, and Nojun Kwak. Music source separation using stacked hourglass networks. *arXiv preprint arXiv:1805.08559*, 2018.

[3] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation, December 2017.

[4] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Surrey, UK*, pages 293–305, 2018.

[5] Naoya Takahashi and Yuki Mitsufuji. Multi-scale multi-band densenets for audio source separation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 21–25. IEEE, 2017.

[6] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki Mitsufuji. Improving music source separation based on deep neural networks through data augmentation and network blending. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 261–265. IEEE, 2017.