MDX
Music Demixing Workshop

CWS-PResUNet: Music Source Separation with Channel-wise Subband Phase-aware ResUNet

Haohe Liu[1, 2], Qiuqiang Kong[1], and Jiafeng Liu[1]

1 Sound, Audio, and Music Intelligence (SAMI) Group, ByteDance **2** The Ohio State University

In partnership with

ISMIR2021

Abstract

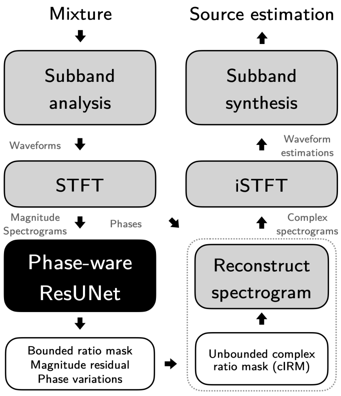
Music source separation (MSS) shows active progress with deep learning models in recent years. Many MSS models perform separations on spectrogram by estimating bounded ratio masks and reusing the phases of the mixture. When using convolutional neural networks (CNN), weights are usually shared within spectrogram during convolution regardless of the different patterns between frequency bands. In this study, we propose a new MSS model, channel-wise subband phase-aware ResUNet (CWS-PResUNet), to decompose signals into subbands and estimate the source unbound complex ideal ratio mask (cIRM) for MSS. CWS-PResUNet utilizes channel-wise subband (CWS) feature to limit unnecessary global weights sharing on the spectrogram and reduce computational resource consumptions. The saved computational cost and memory can in turn allow for a larger architecture. On the MUSDB18HQ test set, we propose a 276-layer CWS-PResUNet and achieve state-of-the-art (SoTA) performance on `vocals` with an 8.92 signal-to-distortion ratio (SDR) score. By combining CWS-PResUNet and Demucs, our ByteMSS system ranks the 2nd on `vocals` score and 5th on average score in the 2021 ISMIR Music Demixing (MDX) Challenge limited training data track (leaderboard A). Our code and pre-trained models are publicly available[1].

Introductions and Related Works

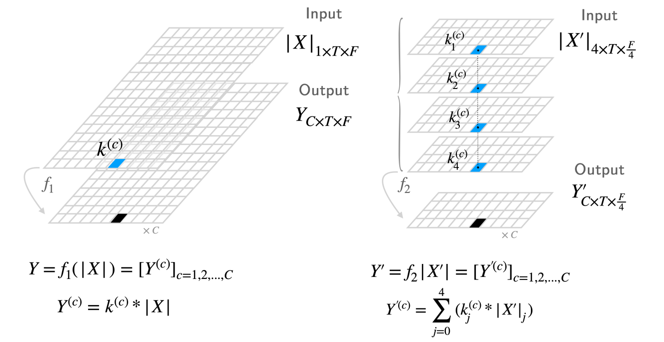
Music source separation aims at decomposing a music mixture into several soundtracks, such as `Vocals`, `Bass`, `Drums`, and `Other` tracks. It is closely related to topics like music transcription, remixing, and retrieval. Based on deep learning models, most of the early studies (Jansson et al., 2017; Takahashi et al., 2018) perform separations in the frequency domain by estimating the ideal ratio masks (IRM) of the magnitude spectrogram and reusing the phase of the mixture. Later, time-domain models (Défossez et al., 2019) start to demonstrate SoTA performance using direct waveform modeling, which does not involve transformations like short-time fourier transform (STFT). In this case, phase information can be implicitly estimated and models will not be restricted with the fixed time-frequency resolution. To enhance the MSS performance, Y. Liu et al. (2020) chooses to employ self-attension mechanism and Dense-UNet architecture. Choi et al. (2019) compares the performance of several types of UNet built with different intermediate blocks. To alleviate the computational cost, Kadandale et al. (2020) designs a multi-task model to replace source-dedicated models. Also, H. Liu et al. (2020) proposes to use the channel-wise subband feature to reduce resource consumptions and improve separation performance. Recently, Kong et al. (2021) conducts experiment on the MSS system theoretical upper bound, which proves the limitation of IRMs and the importance of phase estimation.

In the next section, we will introduce the detailed architecture of CWS-PResUNet as well as ByteMSS, the system we submitted for the MDX Challenge (Mitsufuji et al., 2021).

[1] Open sourced at: https://github.com/haoheliu/2021-ISMIR-MSS-Challenge-CWS-PResUNet



(a) Diagram of CWS-PResUNet.



(b) The comparison of the first 2-D convolution layer of the ResUNet using magnitude (left) and 4-subband CWS (right) spectrogram as the input feature. k and C stand for convolutional kernels and total output channels.

Figure 1: Overview of the CWS-PResUNet and a comparison between using magnitude spectrogram and channel-wise subband spectrogram as the input feature.[2]

Method

CWS-PResUNet is a ResUNet (H. Liu et al., 2021; Zhang et al., 2018) based model integrating the CWS feature (H. Liu et al., 2020) and the cIRM estimation strategies described in Kong et al. (2021). The overall pipeline is summarized in Figure 1a. For a stereo mixture signal $x \in R^{2\times L}$, where L stands for signal length, we first utilize pre-defined analysis filters $h^{(j)}, j = 1, 2, 3, 4$ to perform subband decompositions:

$$x'_{8\times \frac{L}{4}} = [\mathrm{DS}_4(x_{2\times 1\times L} * h^{(j)}_{1\times 64})]_{j=1,2,3,4},$$

where $DS_4(\cdot)$, $*$, and $[\cdot]$ denote the downsampling by 4, convolution, and stacking operators, respectively. Then we calculate the STFT of the downsampled subband signals x' to obtain their magnitude spectrograms $|X'|_{8\times T\times \frac{F}{4}}$.

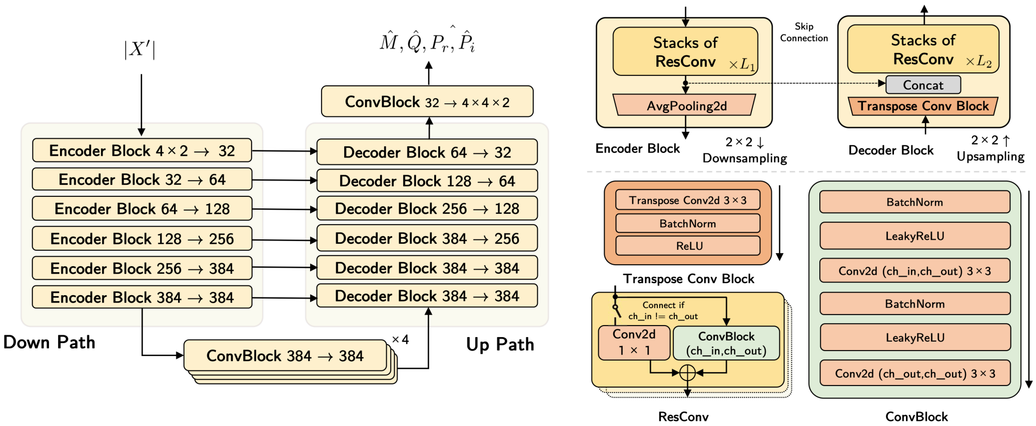


Figure 2: The architecture of Phase-aware ResUNet

As is shown in Figure 2, the phase-aware ResUNet is a symmetric architecture containing a down-sampling and an up-sampling path with skip-connections between the same level. It accepts $|X'|$ as input and estimates four tensors with the same shape: mask estimation $\hat{M}$,

[2]We use mono signal for simple illustration.

phase variation $\hat{P}_r$, $\hat{P}_i$, and direct magnitude prediction $\hat{Q}$. The complex spectrogram can be reconstructed with the following equation:

$$\hat{S}' = \text{relu}(|X'| \odot \text{sigmoid}(\hat{M}) + \hat{Q})e^{j(\angle X' + \angle \hat{\theta})},$$

in which $cos\angle\hat{\theta} = \hat{P}_r/(\sqrt{\hat{P}_r^2 + \hat{P}_i^2})$ and $sin\angle\hat{\theta} = \hat{P}_i/(\sqrt{\hat{P}_r^2 + \hat{P}_i^2})$. We pass the mask estimation $\hat{M}$ through a sigmoid function to obtain a mask with values between 0 and 1. Then by estimating $\hat{Q}$ and $\hat{\theta}$, models can avoid estimating mask with only bounded values and using mixture phase to calculate the unbounded cIRM. We use relu activation to ensure the positve magnitude value. Finally, after the inverse STFT, we perform subband reconstructions to obtain the source estimation $\hat{s}$:

$$\hat{s}_{2\times L} = \sum_{j=1}^{4} (\text{US}_4(\hat{s}'_{2\times 4\times \frac{L}{4}}) * g^{(j)}_{4\times 64}),$$

where $g^{(j)}, j = 1, 2, 3, 4$ are the pre-defined synthesis filters and $US_4(\cdot)$ is the zero-insertion upsampling function.

As is illustrated in Figure 1b, the CWS feature can make the CNN feature-maps smaller and save computational resources. Also, models become more efficient by enlarging receptive fields and diverging subband information into different channels. Our model for `vocals` is optimized by calculating L1 loss between $\hat{s}$ and its target source s. Despite we also use a model dedicated to separating the `other` track, we notice estimating and optimizing four sources together can result in a 0.2 SDR (Vincent et al., 2006) gain on `other`. So, we calculate both L1 loss and energy-conservation loss across four sources to optimize the model for `other`. Our CWS-PResUNet models for `bass` and `drums` reported in the next section employ the same setup as the model for `other`.

In our ByteMSS system, we set up the open-sourced Demucs (Défossez et al., 2019) to separate `bass` and `drums` tracks because it performs better than CWS-PResUNet on these two sources. Demucs is a time-domain MSS model. In our study, we adopted the open-sourced pre-trained Demucs[3] and do not apply the shift trick because it will slow down the inference speed. Also, we utilize a 276-layer CWS-PResUNet to separate the `vocals` track and a 166-layer CWS-PResUNet for the `other` track. We set the latter with fewer layers for faster inference speed concerns.

Experiments

Our models are optimized using the training subset of MUSDB18HQ (Rafii et al., 2019). We calculate the STFT of the downsampled 11.05 kHz subband signals with a window length of 512 and a window shift of 110. We use Adam optimizer with an initial learning rate of 0.001 and exponential decay. CWS-PResUNet takes approximately four days to train on a Tesla V100 GPU. During inference, we utilize a 10-second long `boxcar` windowing function (Schuster et al., 2008) with no overlapping to segment the signal. For evaluation, we report the SDR on the MUSDB18HQ test set with the open-sourced *museval* tool (Stöter et al., 2018).

The subband analysis and synthesis operations usually cannot achieve perfect reconstruction. To assess the errors introduced by subband operations, we decompose the test set `vocals` tracks into 2,4, and 8 subbands and reconstruct them back for evaluations. As is presented in Table 1, in all cases subband reconstructions achieve high performance with only neglectable errors.

[3]https://github.com/facebookresearch/demucs

Table 1. The subband reconstruction performance. Evaluated on MUSDB18HQ test set vocals tracks.

Subband numbers	2	4	8
SDR	102.3	93.7	79.9

Table 2 lists the results of the baselines and our proposed systems. Our CWS-PResUNets achieve an SDR of 8.92 and 5.84 on `vocals` and `other` sources, respectively, outperforming the baseline X-UMX (Sawata et al., 2021), D3Net (Takahashi & Mitsufuji, 2020), and Demucs systems by a large margin. Demucs performs better than CWS-PResUNet on `bass` and `drums` tracks. We assume that is because time-domain models can learn better representations than time-frequency features so are more suitable for separating percussive and band-limited sources. The average performance of our ByteMSS system is 6.97, marking a SoTA performance on MSS. Considering the high performance of the `vocals` model, we also attempt to separate three instrumental sources from `mixture` minus `vocals`. In this case, the average score remains 6.97, in which the `drums` score increase to 6.72 but the other three sources drop slightly. In the future, we will address the integration of time and frequency models for the compensations in both domains.

Table 2. Evaluation results on MUSDB18HQ test set.

Models	Vocals	Drums	Bass	Other	Average
X-UMX	6.61	6.47	5.43	4.64	5.79
D3Net	7.24	7.01	5.25	4.53	6.01
Demucs	6.89	**6.57**	**6.53**	5.14	6.28
CWS-PResUNet	**8.92**	6.38	5.93	**5.84**	6.77
ByteMSS	8.92	6.57	6.53	5.84	**6.97**

Conclusions

Our experiment result shows CWS-PResUNet can achieve a leading performance on the separation of `vocals` and `other` tracks. And channel-wise subband feature is an effective alternative to magnitude spectrogram on music source separation task.

Acknowledgements

This project is funded by ByteDance Inc. We acknowledge the supports from Haonan Chen for testing our system.

Reference

Choi, W., Kim, M., Chung, J., Lee, D., & Jung, S. (2019). Investigating u-nets with various intermediate blocks for spectrogram-based singing voice separation. *arXiv Preprint arXiv:1912.02591*.

Défossez, A., Usunier, N., Bottou, L., & Bach, F. (2019). Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv Preprint arXiv:1909.01174*.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). *Singing voice separation with deep u-net convolutional networks*.

Kadandale, V. S., Montesinos, J. F., Haro, G., & Gómez, E. (2020). *Multi-task u-net for music source separation*.

Kong, Q., Cao, Y., Liu, H., Choi, K., & Wang, Y. (2021). Decoupling magnitude and phase estimation with deep ResUNet for music source separation. *arXiv Preprint arXiv:2109.05418*.

Liu, H., Kong, Q., Tian, Q., Zhao, Y., Wang, D., Huang, C., & Wang, Y. (2021). VoiceFixer: Toward general speech restoration with neural vocoder. *arXiv Preprint arXiv:2109.13731*.

Liu, H., Xie, L., Wu, J., & Yang, G. (2020). Channel-wise subband input for better voice and accompaniment separation on high resolution music. *arXiv Preprint arXiv:2008.05216*.

Liu, Y., Thoshkahna, B., Milani, A., & Kristjansson, T. (2020). Voice and accompaniment separation in music using self-attention convolutional neural network. *arXiv Preprint arXiv:2003.08954*.

Mitsufuji, Y., Fabbro, G., Uhlich, S., & Stöter, F.-R. (2021). Music demixing challenge at ISMIR 2021. *arXiv Preprint arXiv:2108.13559*.

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. (2019). *MUSDB18-HQ-an uncompressed version of MUSDB18*.

Sawata, R., Uhlich, S., Takahashi, S., & Mitsufuji, Y. (2021). All for one and one for all: Improving music separation by bridging networks. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 51–55.

Schuster, S., Scheiblhofer, S., & Stelzer, A. (2008). The influence of windowing on bias and variance of DFT-based frequency and phase estimation. *IEEE Transactions on Instrumentation and Measurement*, 1975–1990.

Stöter, F.-R., Liutkus, A., & Ito, N. (2018). *The signal separation evaluation campaign*. 293–305.

Takahashi, N., Goswami, N., & Mitsufuji, Y. (2018). MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation. *International Workshop on Acoustic Signal Enhancement*, 106–110.

Takahashi, N., & Mitsufuji, Y. (2020). D3net: Densely connected multidilated densenet for music source separation. *arXiv Preprint arXiv:2010.01733*.

Vincent, E., Gribonval, R., & Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 1462–1469.

Zhang, Z., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 749–753.