Investigating U-Nets with various Intermediate Blocks for Spectrogram-based Singing Voice Separation

Woosung Choi [1] Minseok Kim [1] Jaehwa Chung [2] Daewon Lee [3] Soonyoung Jung [1]

1 Department of Computer Science and Engineering, Korea University, Republic of Korea

2 Department of Computer Science, Korea National Open University, Republic of Korea

3 Department of Computer Engineering, Seokyeong University, Republic of Korea

INTRODUCTION

Singing Voice Separation (SVS)

- separate singing voice from a given mixed musical signal
  - original track: Paul Kim - Empty ***(click)***
  - Demo: U-Net with TFC-TDF ***(click)***
- Related Works: roughly categorized into two groups
  - waveform-to-waveform models: trie to generate the vocal waveforms directly
  - **spectrogram-based models**: estimate spectrograms (usually magnitude) of vocal waveforms
    1. Apply Short-Time Fourier Transform (STFT) on a mixture waveform to obtain the input spectrograms.
    2. Estimate the vocal spectrograms based on these inputs
    3. Restore the vocal waveform with inverse STFT (iSTFT).

Related Works: U-Net-based SVS Models

- U-Net: an encoder-decoder structure with symmetric skip connections
  - These symmetric skip connections allow models to recover fine-grained details of the target object during decoding effectively.
- U-Net-like Models for SVS (or MSS)
  - They have revealed that U-Net-like architectures can provide promising performance for SVS and MSS

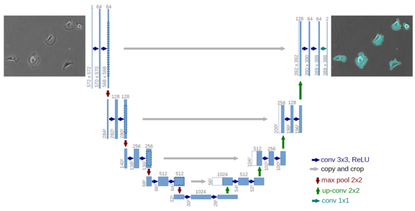

Figure 1. The U-Net Structure

Related Works: Intermediate blocks of U-Nets

- Existing works proposed various types of neural networks for intermediate blocks.
  - Some models used simple **Convolutional Neural Networks (CNNs)**
  - Other advanced models tried more **complex intermediate blocks**.
    - MMDenseLSTM
      - Takahashi, Naoya, Nabarun Goswami, and Yuki Mitsufuji. "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation." 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, 2018.
- (Motivation) No existing works that evaluate and directly compare these different types of blocks

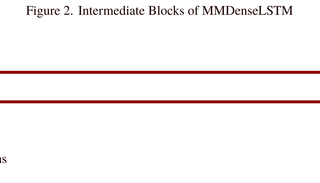

Figure 2. Intermediate Blocks of MMDenseLSTM

METHODS

The scope of this paper

1. We designed several types of blocks based on different design strategies
2. For each type of block, we implemented at least one SVS model, which are all based on an identical U-Net framework for fair comparisons
3. We summarize the experimental results and discuss the effect of each design choice

U-Net-based SVS Framework (shared by every models)

1. Complex as Channel Framework
   - a singing voice separation framework based on complex-valued spectrogram estimation
   - It takes a c-channeled mixture signal, and outputs c-channeled singing voice signal
2. U-Net Architecture for Spectrogram Estimation
   - A Generalized U-Net for SVS
   - two main components: (1) intermediate blocks (2) up/down sampling layers

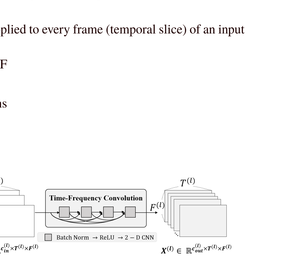

Figure 3. Complex as Channel Framework (CaC)

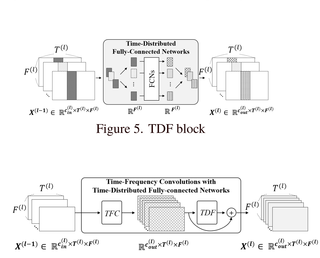

Figure 4. U-Net for Complex-valued Spectrogram Estimation

Our Intermediate blocks

1. Time-Distributed Blocks: do not have any inter-frame operations. They are applied to every frame (temporal slice) of an input
   - Time-Distributed Fully-connected networks (TDF): Figure 5
   - Time-Distributed Convolutions (TDC): a 1-d convolutional version of TDF
2. Time-Frequency Blocks: consider both the time and frequency dimension
   - Time-Frequency Convolutions (TFC): Densly-connected 2-D Convolutions
   - Time-Frequency Convolutions with TDF (TFC-TDF): TFC with TDF
   - Time-Distributed Convolutions with RNNs (TDC-RNN): TDF wit RNN

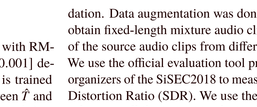

Figure 5. TDF block

Figure 6. TDC block

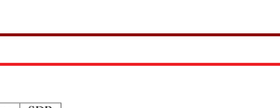

Figure 7. TFC block

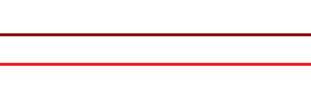

Figure 8. TFC-TDF block

Experiment Environment

1 Dataset

Train and test data were obtained from the MUSDB dataset (Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, Rachel Bittner. MUSDB18 - a corpus for music separation. 2017). The train and test sets of MUSDB have 100 and 50 musical tracks each, all stereo and sampled at 44100 Hz.

2 Model Configurations

We set $c_{in}^{(1)}$, the number of internal channels to be 24. Each model uses a single type of block for its intermediate blocks. We usually used an FFT window size of 2048 and a hop size of 1024 for STFT. However, we used a larger window size in some models for a fair comparison with SOTA methods.

3 Training and Evaluation

Weights of each model were optimized with RMSprop with learning rate $lr \in [0.0005, 0.001]$ depending on model depth. Each model is trained to minimize the mean square error between $\hat{\boldsymbol{T}}$ and $\mathbf{T}$. We use the default validation set (14 tracks) as defined in the *MUSDB* package, and use the Mean Squared Error (MSE) between target and estimated signal (waveform) as the validation metric for validation. Data augmentation was done on the fly to obtain fixed-length mixture audio clips comprised of the source audio clips from different tracks. We use the official evaluation tool provided by the organizers of the SiSEC2018 to measure Source-to-Distortion Ratio (SDR). We use the median SDR value over all the test set tracks to obtain the overall SDR performance for each run, as done in the SiSEC2018. We report the average of 'median SDR values' over three runs for each model.

RESULTS

Experimental Results

block type	# blocks	# params	SDR
TDC (w/ sampling)	17	0.54M	**4.86**
TDC (w/o sampling)	17	0.52M	3.78
TDC (w/o sampling)	3	0.09M	3.56
TDF (w/o hidden layer)	17	2.83M	4.75
TDF (w/ hidden layer)	17	1.44M	4.05
TDF (w/ hidden layer)	3	1.19M	4.01

Table 1. Evaluation results of Time-Distributed Blocks.

model	# parameters	SDR (vocals)
DGRU-DGConv	more than 1.9M	6.99
TAK1	1.22M	6.60
UMX	8.89M	6.32
TFC-TDF (small)	0.99M	**7.07** ±.08
TFC-TDF (large)	2.24M	**7.98** ±.07

Table 3. Comparison: SDR median value on test set.

model	sampling	# blocks	# params	SDR
TFC	O	17	1.56M	6.89
TFC	X	17	1.56M	6.75
TDC-RNN	O	17	2.08M	6.69
TFC-TDF	O	7	0.99M	7.07
TFC-TDF	O	17	1.93M	**7.12**

Table 2. Evaluation results of Time-Frequency Blocks.

esimation	n_fft	# blocks	# params	SDR
CaC	2048	7	0.99M	**7.07**
Mag	2048	7	0.99M	6.43
CaC	4096	9	2.24M	**7.98**
Mag	4096	9	2.24M	7.24

Table 4. Comparison of TFC-TDFs: CaC vs Mag

Audio Samples

- original track: Al James - Schoolboy Fascination ***(click)***
- Separated Results ***(click)***

DISCUSSION

Our work provides a practical guideline for choosing fundamental building blocks to develop an SVS or MSS model based on the U-Net architecture as follows.

- TDC-based models are sensitive to the number of blocks, compared to TDF-based models.
- Using down/up-sampling is important for CNN-based blocks, especially in the frequency dimension.
- Stacking 2-D CNNs is a simple but effective way to capture T and F features, compared to TDC-RNNs.
- Injecting a time-distributed block to a time-frequency block can improve SDR.
- A simple extension from a magnitude-only U-Net to a CaC U-Net can improve SDR.

Our work is not limited to the U-Net-architecture nor MSS. Blocks can be used as core components in more complex architectures as well. We can use different types of blocks for a single model, meaning that a lot of space remains for improvement. Also, our observations can be exploited in other MIR tasks such as Automatic Music Transcription (AMT) or Music Generation: for example, we expect that injecting TDFs to intermediate blocks for f_0 estimation model can improve performance since fully-connected layer can efficiently model long-range correlations such as harmonics.

References

[1] UNOFFICIAL Template for Poster with Radboud University, Donders Institute layout https://ko.overleaf.com/latex/templates/template-ru-di-poster/rtmwgwfyhhsv

[2] Please check the refernce section of the original paper

Acknowledgments: This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2019R1F1A1062719, NRF-2020R1A2C1012624).

e-mail: jsy@korea.ac.kr

KOREA UNIVERSITY

Presented at the <ISMIR 2020> in <Virtual Conference> on Monday 12th October 2020