Complex Ratio Masking for Singing Voice Separation

Yixuan Zhang[1], Yuzhou Liu[1] and DeLiang Wang[1,2]

1. Department of Computer Science & Engineering, The Ohio State University

2. Center of Cognitive and Brain Sciences, The Ohio State University

Poster Number: 2587

Summary

Music source separation is important for applications such as karaoke and remixing. Much of previous research focuses on estimating magnitude short-time Fourier transform (STFT) and discarding phase information. We observe that, for singing voice separation, phase has the potential to make considerable improvement in separation quality. This paper proposes a complex-domain deep learning method for voice and accompaniment separation. The proposed method employs DenseUNet with self attention to estimate the real and imaginary components of STFT for each sound source. A simple ensemble technique is introduced to further improve separation performance. Evaluation results demonstrate that the proposed method outperforms recent state-of-the-art models for both separated voice and accompaniment.

1. Importance of Phase in Singing Voice Separation

For both singing voice and accompaniment, the use of clean phase leads to considerable improvement, about 4 to 5 dB on average.

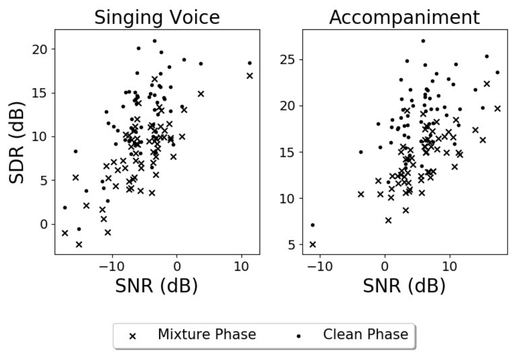

Figure 1. SDR of separated singing voice (left) and accompaniment (right) with mixture phase versus clean phase

SA-DenseUNet [1] is used to estimate magnitude spectrograms of singing voice and accompaniment for 63 songs with different signal-to-noise ratios (SNRs) from the test set described in Experiments section and the comparison between the signal-to-distortion ratio (SDR) of output audios re-synthesized with clean phase versus mixture phase is shown in Figure 1.

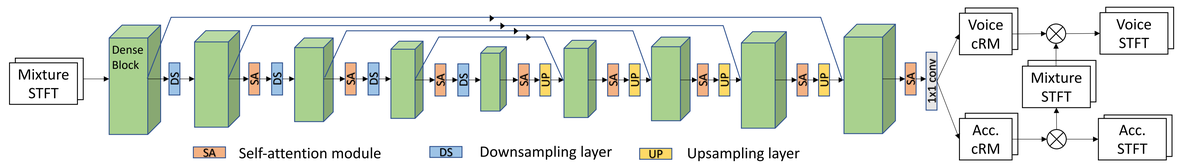

Figure 2. Diagram of network structure. The input, masks, outputs are all defined in the complex domain. Acc. refers to accompaniment.

2. Methods

Complex SA-DenseUNet

We extend SA-DenseUNet [1] to estimate the real and imaginary components of the complex ideal ratio mask (cIRM) [2] of each source, which are then multiplied with the complex STFT of the mixture audio to get the estimated complex STFT of each source. The loss function is defined as,

$$L = \sum_{j=1,2} [\left|Re(S_j - cRM_j \odot Y)\right| + \left|Im(S_j - cRM_j \odot Y)\right|]$$

where cRM_j is an estimate of the cIRM for source j, and $\odot$ denotes element-wise multiplication. Y denotes the complex STFT of the input mixture, and S_1 and S_2 represent the complex STFT of singing voice and accompaniment respectively.

Multi-context Averaging

We introduce a simple ensemble learning technique which exploits different contexts of the input by using different window lengths.

We train 3 complex SA-DenseUNets with different window lengths (32ms, 64ms and 128ms), which results in different context windows, with lengths ranging from 5s to 20s. The network is thus enabled to model music repetition at different levels.

The outputs of all 3 sub-networks are transformed to the time domain by inverse STFT and then averaged to get the final waveform.

3. Experiments

Training Set

- The training set contains 50 songs from DSD100's Dev Set and 400 songs generated by randomly scaling, shifting, remixing different music sources from 50 songs for data augmentation

Evaluation Set

- The test set contains 63 songs from DSD100 Test Set, MedleyDB and CCMixter datasets.

Table 1 Experimental results on the test set

	Singing Voice			Accompaniment		
Metric (dB)	SDR	SIR	SAR	SDR	SIR	SAR
SA-DenseUNet [1]	8.08	15.44	9.34	14.10	18.42	16.50
Complex SA-DenseUNet	9.09	20.15	9.76	15.21	22.77	16.45
Multi-context Averaging	9.73	20.76	10.36	15.57	21.78	17.28

Complex SA-DenseUNet improves mean SDRs of singing voice and accompaniment by 1.01 dB and 1.11 dB. Multi-context averaging further improves the SDR of singing voice and accompaniment by 0.64 dB and 0.36 dB respectively.

4. Comparison with other methods

Table 2 Comparison of median SDR values on DSD100 datasets

Metric (dB)	Singing Voice	Accompaniment
MMDenseNet [3]	6.00	12.10
MMDenseLSTM [4]	6.31	12.73
SA-SHN-4 [5]	6.44	12.60
SA-DenseUNet [1]	7.72	13.90
Proposed	**9.78**	**15.20**

- The evaluation and comparisons are conducted on the DSD100 Test Set.
- A median SDR score is the median of SDR scores of all songs, and it is documented separately for singing voice and accompaniment.
- Our method outperforms the strongest baseline of SA-DenseUNet by 2.06 dB for singing voice and 1.30 dB for accompaniment.

5. Conclusions

- We observe that phase is important for singing voice separation, and the cIRM is an effective training target where the loss is defined in terms of complex spectrogram.
- A simple ensemble learning technique is found to be effective for singing voice separation.
- Evaluation results show that the proposed method outperforms other state-of-the-art methods.

Selected references

1. Y. Liu, B. Thoshkahna, A. Milani, and T. Kristjansson, "Voice and accompaniment separation in music using self-attention convolutional neural network," arXiv preprint arXiv:2003.08954, 2020.

2. D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, pp. 483–492, 2016.

3. N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in Proc. WASPAA, 2017, pp. 21–25.

4. N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in Proc. 16th IEEE International Workshop on Acoustic Signal Enhancement, 2018, pp. 106–110.

5. W. Yuan, S. Wang, X. Li, M. Unoki, and W. Wang, "A skip attention mechanism for monaural singing voice separation," IEEE Signal Processing Letters, vol. 26, pp. 1481–1485, 2019.