Audionamix

Online Singing Voice Separation using a Recurrent One-Dimensional U-Net Trained with Deep Feature Losses

Clément Doire
Audionamix, France

Abstract

We propose an online approach to the singing voice separation problem. Based on a combination of one-dimensional convolutional layers along the frequency axis and recurrent layers to enforce temporal coherency, state-of-the-art performance is achieved. The concept of using deep features in the loss function to guide training and improve the model's performance is also investigated.

Problem Statement

Let the singing voice, background music, and input mixture stereo magnitude spectrograms at time-frame t be given by $\mathbf{V}_t$, $\mathbf{B}_t$ and $\mathbf{X}_t$ respectively. We define the task of online singing voice separation as finding the time-varying masks $\mathbf{M}_{V_t}$ and $\mathbf{M}_{B_t}$ such that:

$$\hat{\mathbf{V}}_t = \mathbf{M}_{V_t} \odot \mathbf{X}_t \tag{1}$$

$$\hat{\mathbf{B}}_t = \mathbf{M}_{B_t} \odot \mathbf{X}_t \tag{2}$$

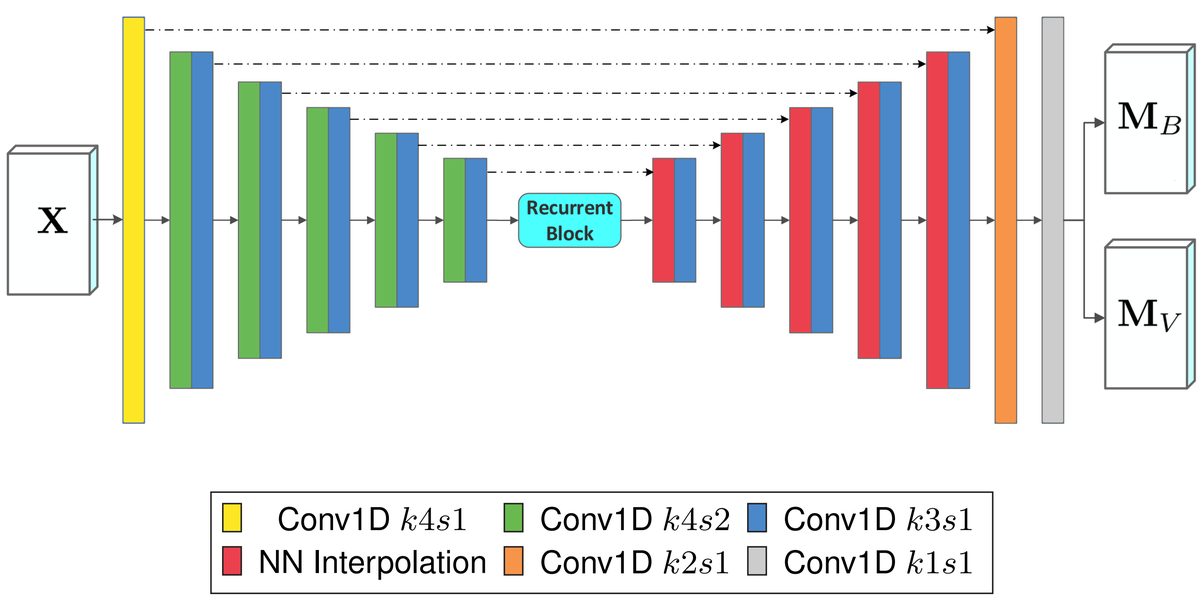
We aim to train a single neural network to compute the two time-varying masks, so that $(\mathbf{M}_{V_t}, \mathbf{M}_{B_t}) = f(\mathbf{X}_t, \mathbf{h}_{t-1})$ where $\mathbf{h}_{t-1}$ contains information about the past time-frames, akin to a hidden state.

In order to avoid defining explicitly what the target masks should be during training, we define the spectrogram reconstruction loss function as [1]

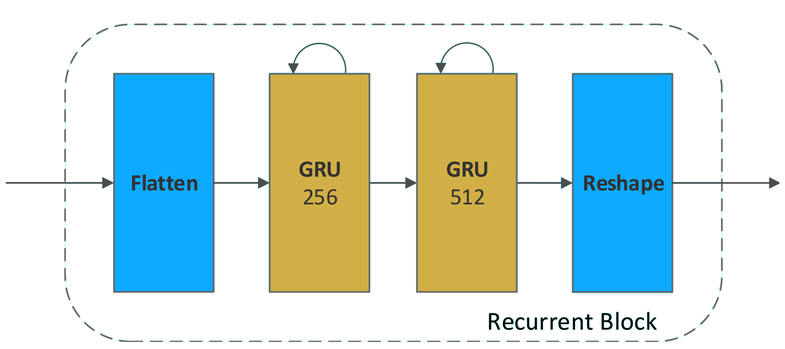
$$\mathcal{L}_R = \frac{1}{T}\sum_{t=1}^{T}\left(\|\mathbf{V}_t - \mathbf{M}_{V_t} \odot \mathbf{X}_t\|_1 + \|\mathbf{B}_t - \mathbf{M}_{B_t} \odot \mathbf{X}_t\|_1\right) \tag{3}$$

Online Recurrent U-Net

A modified U-Net architecture [2] is used, with one-dimensional convolutional layers computing feature maps along the frequency axis only.



To enforce temporal coherency between successive time-frames, recurrent layers are used in the bottleneck.



Training with Deep Feature Losses

We train a separate CNN to be used as a loss network [3], and incorporate the computed deep features in the objective function. Akin to student-teacher models, we use as our loss network a U-Net [2] trained on the same source separation task. The 2-D convolutions of the loss network can thus steer the GRU layers of the OR-U-Net towards learning to respect meaningful temporal patterns at different scales.

Let $\phi_i(\mathbf{Z})$ be the feature maps at the output of the i^{th} layer of the loss network for an input $\mathbf{Z}$. We define the feature loss at the i^{th} layer as

$$\mathcal{L}_{\phi_i} = \left\|\phi_i(\mathbf{V}) - \phi_i(\hat{\mathbf{V}})\right\|_1 + \left\|\phi_i(\mathbf{B}) - \phi_i(\hat{\mathbf{B}})\right\|_1 \tag{4}$$

where the L_1 norm is computed as the mean absolute value of the array elements. The total loss used during training is therefore

$$\mathcal{L} = w_0\mathcal{L}_R + \sum_{i=1}^{N} w_i\mathcal{L}_{\phi_i} \tag{5}$$

with N the total number of deep features to use in the loss function.

Evaluation

Four methods are compared on the singing voice separation task:

- **LSTM** A network architecture inspired by [4] with 3 LSTM layers of 256 units and a final fully-connected layer, serving as a baseline method.
- **U-Net** The adapted U-Net architecture [2] used as a feature loss network. Does not treat the task in an online manner.
- **OR-U-Net** The proposed Online Recurrent U-Net trained using the L_1 cost function (3).
- **OR-U-Net**$_{df}$ The proposed Online Recurrent U-Net trained using the deep feature losses cost function (5) with $N = 4$ deep features and the decreasing weight strategy $w_i = N - i + 1$, normalized so that $\sum_{i=0}^{N} w_i = 1$.

To train all models, we used the 50 songs of the DSD100 training set, 50 songs from the MedleyDB database and 30 songs from the CCMixter database. Out of the total 130 songs, 20 were randomly assigned to the validation set. All methods are evaluated on the 50 songs of the test set of DSD100, using the masks directly at the output of each network without any post-processing applied.

	Vocals			Background		
Method	SDR	SIR	SAR	SDR	SIR	SAR
LSTM	2.83	6.89	6.02	9.48	12.39	12.99
U-Net	3.21	8.34	5.86	9.81	**13.42**	12.98
OR-U-Net	3.14	7.41	**6.20**	**9.87**	13.25	12.99
OR-U-Net$_{df}$	**3.70**	**9.52**	5.80	9.65	11.84	**14.16**

The proposed OR-U-Net model trained without deep feature losses achieves performance on a par with the U-Net and stronger performance than the baseline LSTM online model on all criteria. The OR-U-Net$_{df}$ model achieves best SDR and SIR overall performance on the vocals separation task, with higher scores than both the base OR-U-Net model and the U-Net loss network used for its training.

References

[1] S. I. Mimilakis, K. Drossos, J. Santos, G. Schuller, T. Virtanen, and Y. Bengio, "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask," *arXiv preprint arXiv:1711.01437v2*, 2018.

[2] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," *18^{th} International Society for Music Information Retrieval (ISMIR) Conference*, 2017.

[3] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," *IEEE International Conference on Computer Vision (ICCV)*, 2016.

[4] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.