Weakly Informed Audio Source Separation

Kilian Schulze-Forster,[1] Clément Doire,[2] Gaël Richard,[1] Roland Badeau[1]
[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
firstname.lastname@telecom-paris.fr
[2] Audionamix, Paris, France

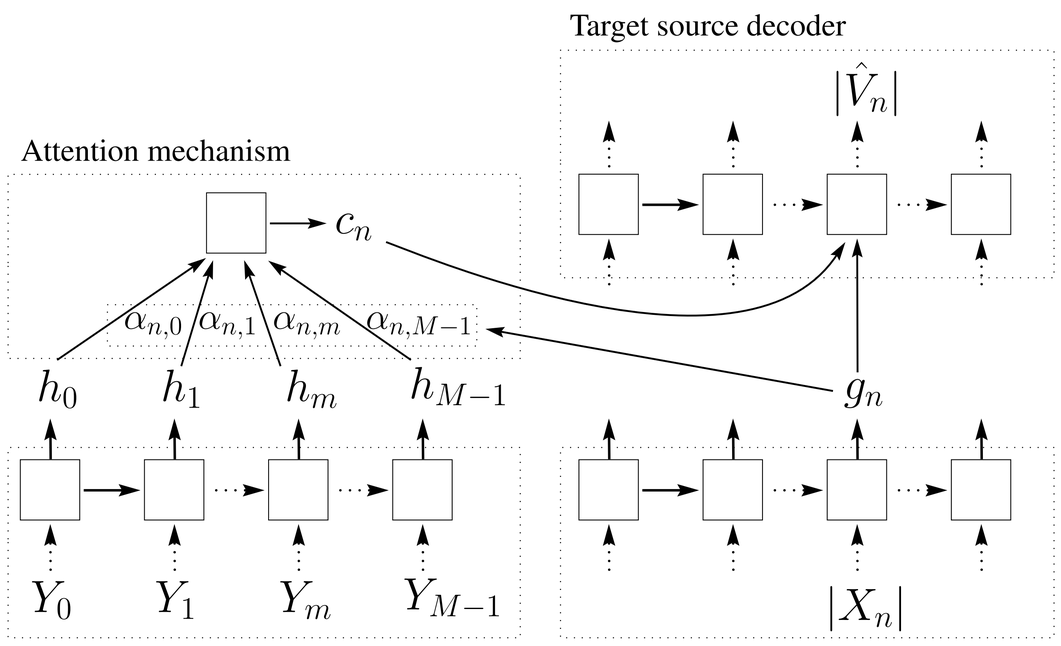
MIP-Frontiers
mip-frontiers.eu

Introduction

- Prior information about the target source can improve audio source separation quality but is usually not available with the necessary level of audio alignment.
- We propose learning to align and separate jointly in order to exploit such weak side information.
- We test the model on a singing voice separation task using artificial side information with different levels of expressiveness.

Proposed Model

- **Inputs:**
  - $|X| \in \mathbb{R}^{F \times N}$: magnitude of mixture STFT with F frequency bands and N time frames
  - $Y \in \mathbb{R}^{D \times M}$: side information with D features and M time steps
- **Output:**
  - $|\hat{V}| \in \mathbb{R}^{F \times N}$: target source magnitude STFT prediction
- An inverse STFT of $|\hat{V}|$ combined with the mixture phase is performed to obtain the target estimation $\hat{v}(t)$ in the time domain.



- The encoders and decoder are bidirectional LSTM-RNNs with two layers.
- The attention mechanism identifies the relevant elements in the side information encoding h for each time step n of the target source decoding and summarizes them in a context vector c_n:

$$s_{n,m} = g_n^\top W h_m \tag{1}$$

$$\alpha_{n,m} = \frac{\exp(s_{n,m})}{\sum_{m=0}^{M-1} \exp(s_{n,m})} \tag{2}$$

$$c_n = \sum_{m=0}^{M-1} h_m \alpha_{n,m} \tag{3}$$

Source Separation Quality Evaluation

- SDR, SAR, SIR are typically computed on non-overlapping frames of one second length [1].
- For frames with a silent target source or prediction, the metrics are undefined.
- Consequently, 45 out of 210 minutes of the MUSDB18 test set are systematically ignored during evaluation.
- This issue has also been observed in [2]
- We propose to evaluate additionally:
  - Predicted Energy at Silence (PES): How much energy is in the prediction when the target source is silent?
  - Energy at Predicted Silence (EPS): How much energy is in the target source when silence is predicted?

Experiments

Can the proposed model align and exploit weak side information?

We perform singing voice separation on the MUSDB18 data set using artificial side information with different levels of expressiveness. All songs are converted to mono, downsampled to 16 kHz, and cut into fragments of 8.2 seconds.

- **Baselines:**
  - **BL1:** Only mixture encoder and target source decoder, no side information
  - **BL2:** Complete architecture, meaningless side information $Y_m = 1$
- **Vocal magnitude as side information:**
  - **M1:** Total magnitude of true vocals for each time frame n.
  - **M2:** Same as M1 but padded with random length at start and end.
- **Vocal activity as side information:**
  - **A1:** Derived from M1, voice active = 1, not active = 0, randomly padded.
  - **A2:** Same as A1, sequences of zeros randomly shortened
  - **A3.1:** Same as A2, also sequences of ones randomly shortened
  - **A3.2:** Same as A3.1, circularly shifted during testing

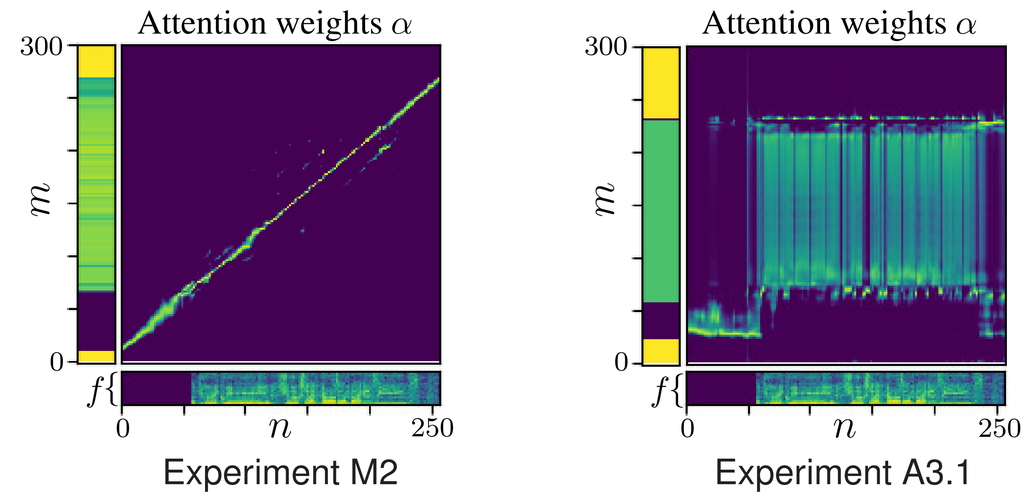
Results

Evaluation scores are given in dB:

	SDR	SAR	SIR	PES	EPS
BL1	2.98	6.41	7.99	-44.89	-25.58
BL2	3.33	6.33	7.78	-43.78	-22.96
M1	3.94	6.28	8.86	-106.15	-46.12
M2	3.89	6.25	9.10	-111.94	-49.75
A1	3.51	6.44	8.00	-85.91	-37.46
A2	3.18	6.25	7.93	-89.87	-36.00
A3.1	3.16	6.17	7.75	-85.20	-34.53
A3.2	3.21	6.30	7.95	-77.64	-32.99

AUDIO EXAMPLES

- The model is able to exploit degraded side information with different length than the mixture STFT for source separation and to align it.
- Vocal magnitude information is useful for both active and non-active vocal frames. All metrics but SAR improve over the baselines.
- Binary vocal activity information is useful mainly to identify silent vocal parts. PES and EPS improve while the other metrics do not change much compared to the baselines
- Below, we present the attention weights α containing alignment information. The side information is shown vertically on the left the true vocals spectrogram below. Lighter color indicates higher values.



References

[1] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito, "The 2018 signal separation evaluation campaign," *International Conference on Latent Variable Analysis and Signal Separation*, pp. 293–305, 2018.

[2] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Jointly detecting and separating singing voice: A multi-task approach," *International Conference on Latent Variable Analysis and Signal Separation*, pp. 329–339, 2018.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068