Weakly Informed Audio Source Separation

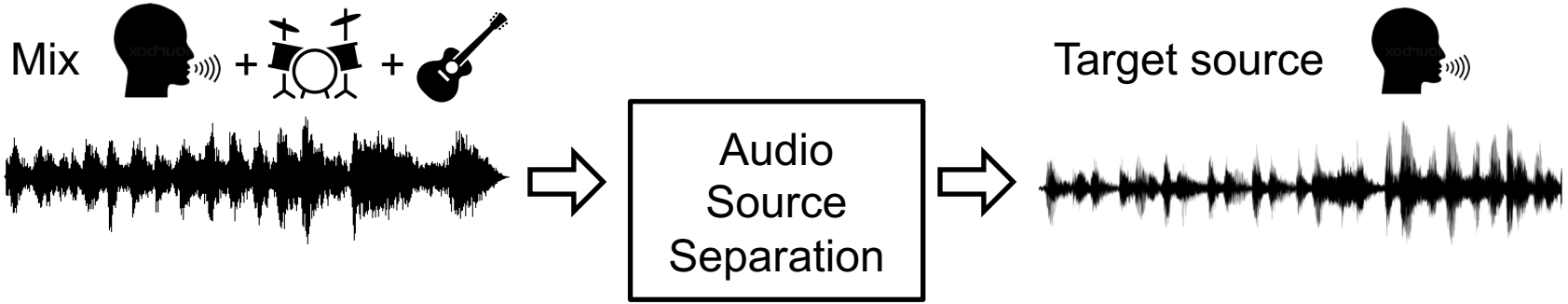
Kilian Schulze-Forster,[1] Clement Doire,[2] Gaël Richard,[1]
Roland Badeau[1]
[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] Audionamix, Paris, France

MIPFrontiers

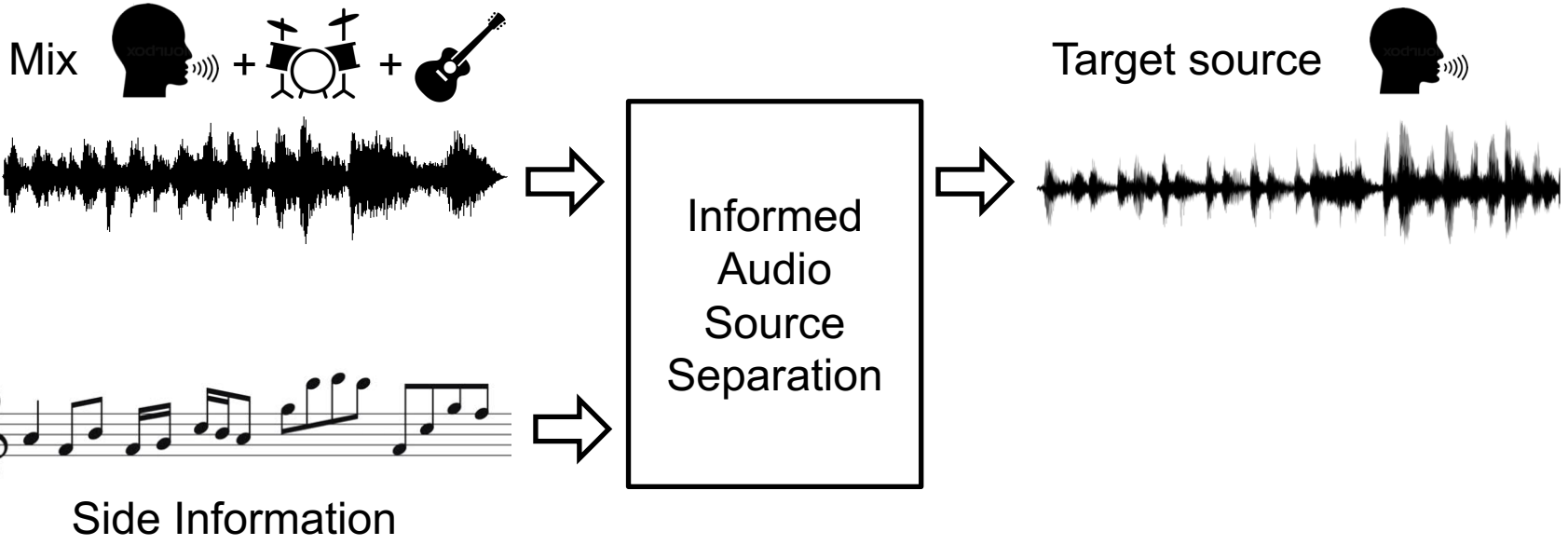
Sound check

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

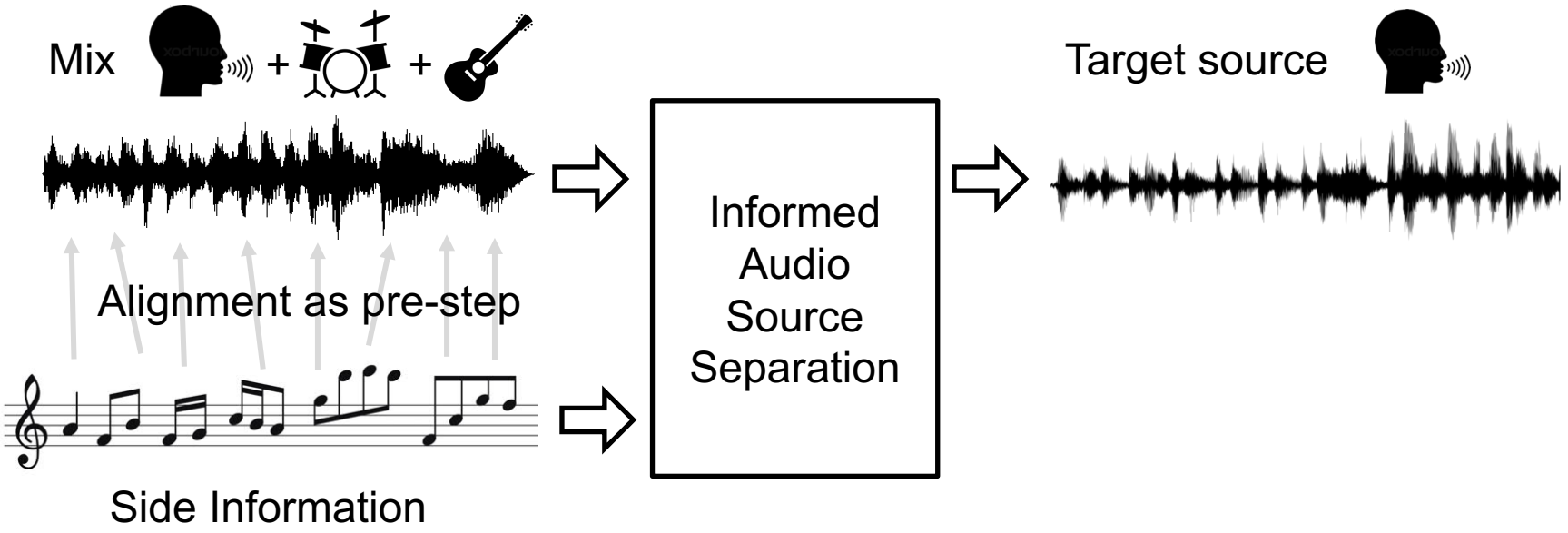
Introduction



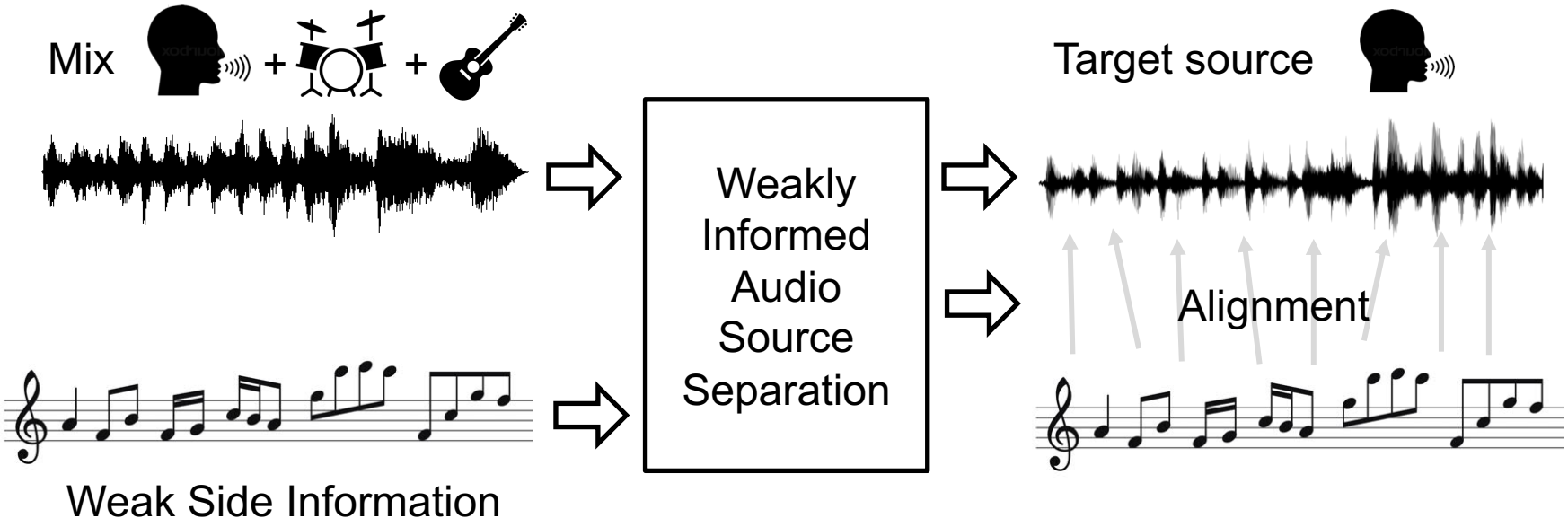
Introduction



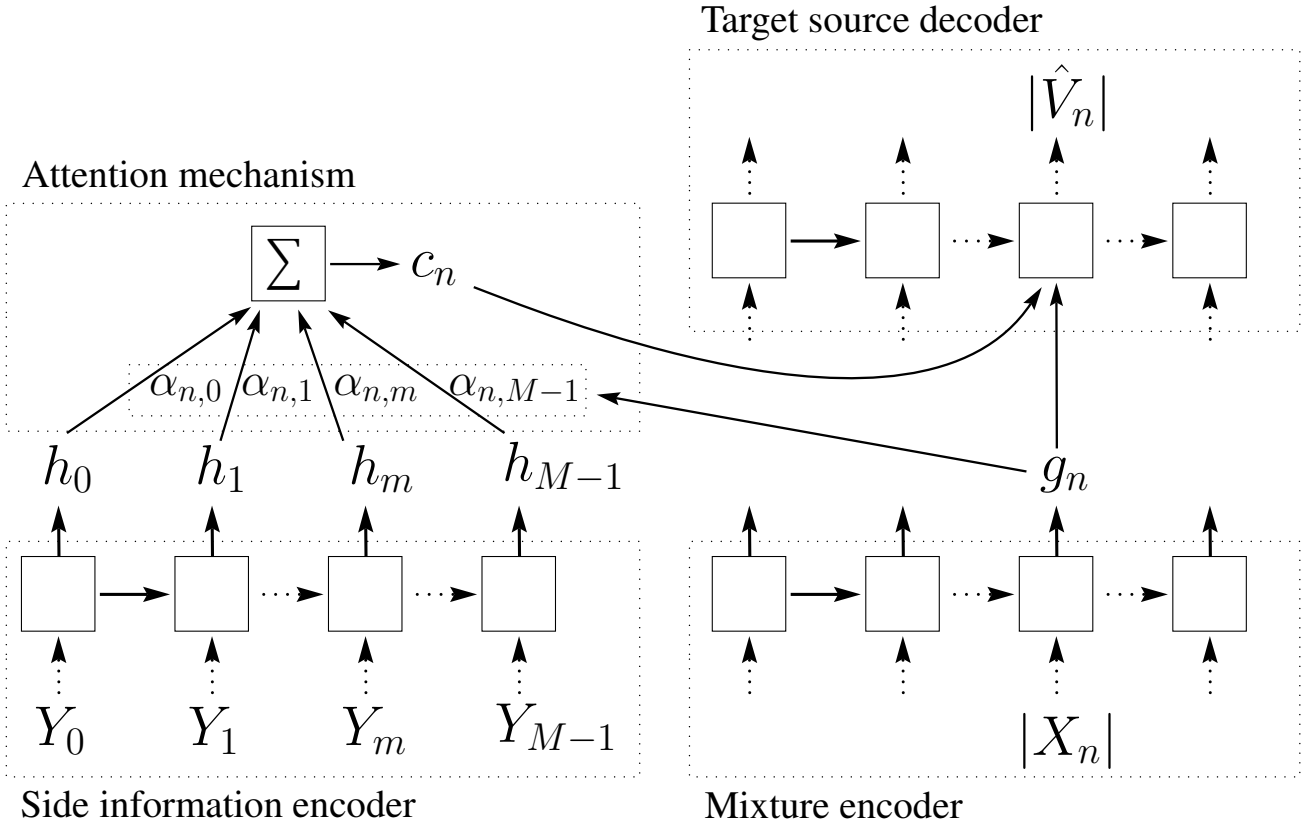
Introduction



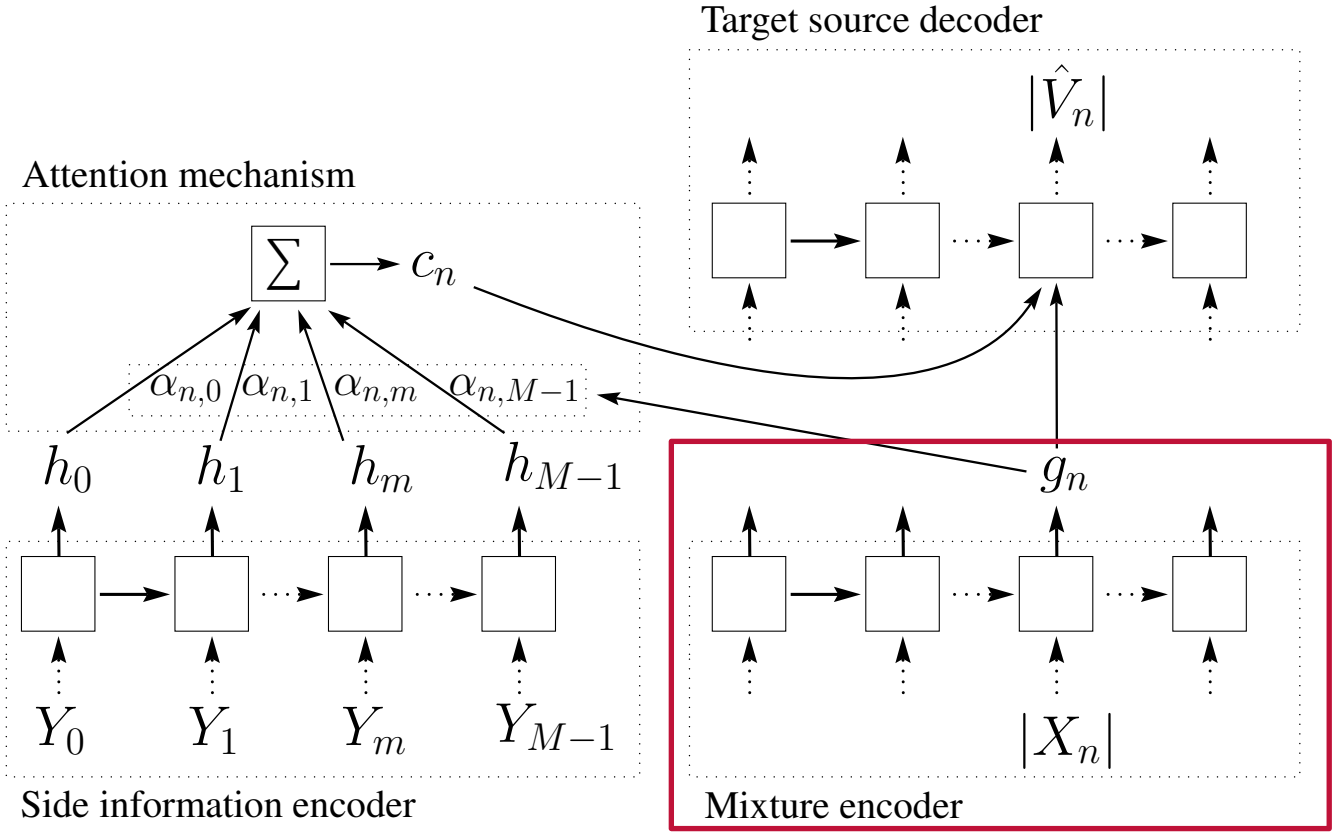
Introduction



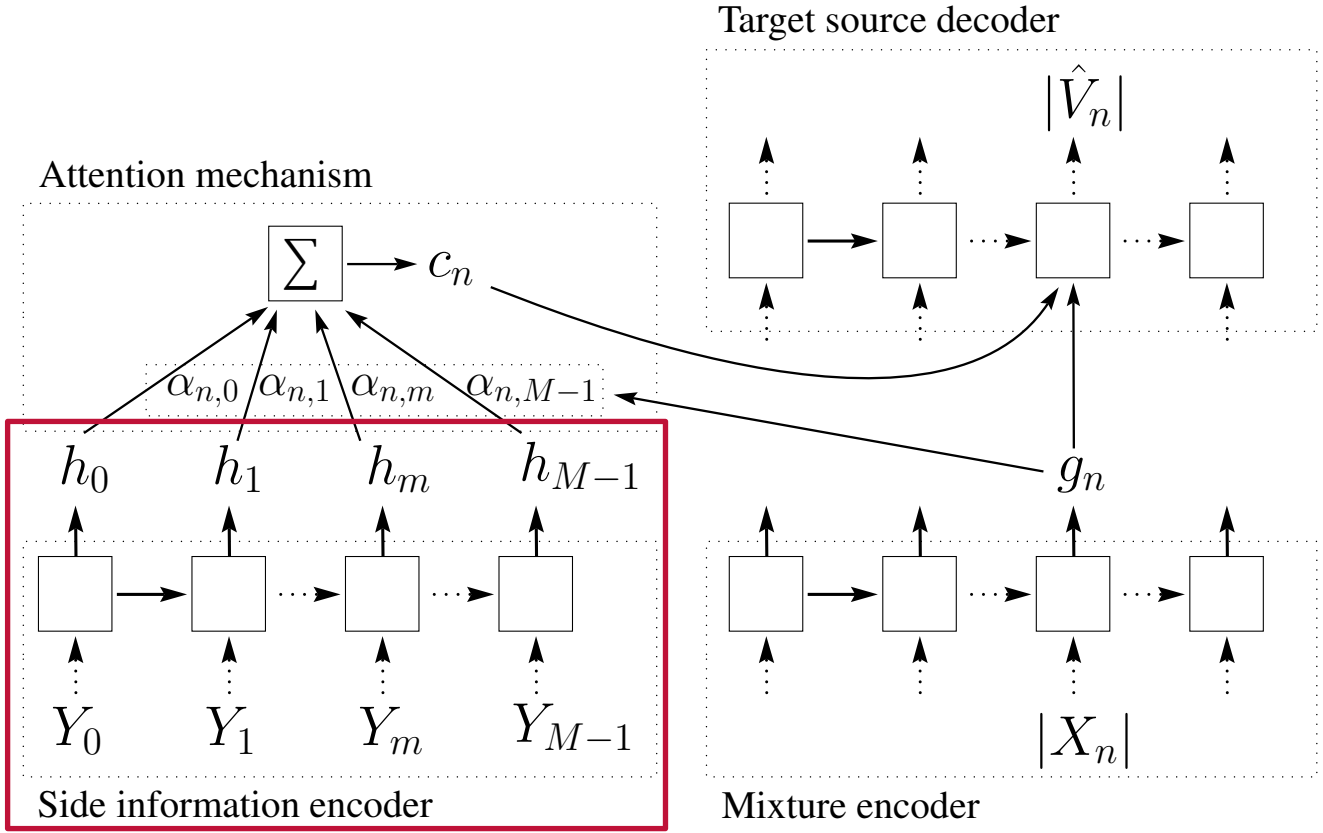
Proposed Model: Learn to Align and Separate Jointly



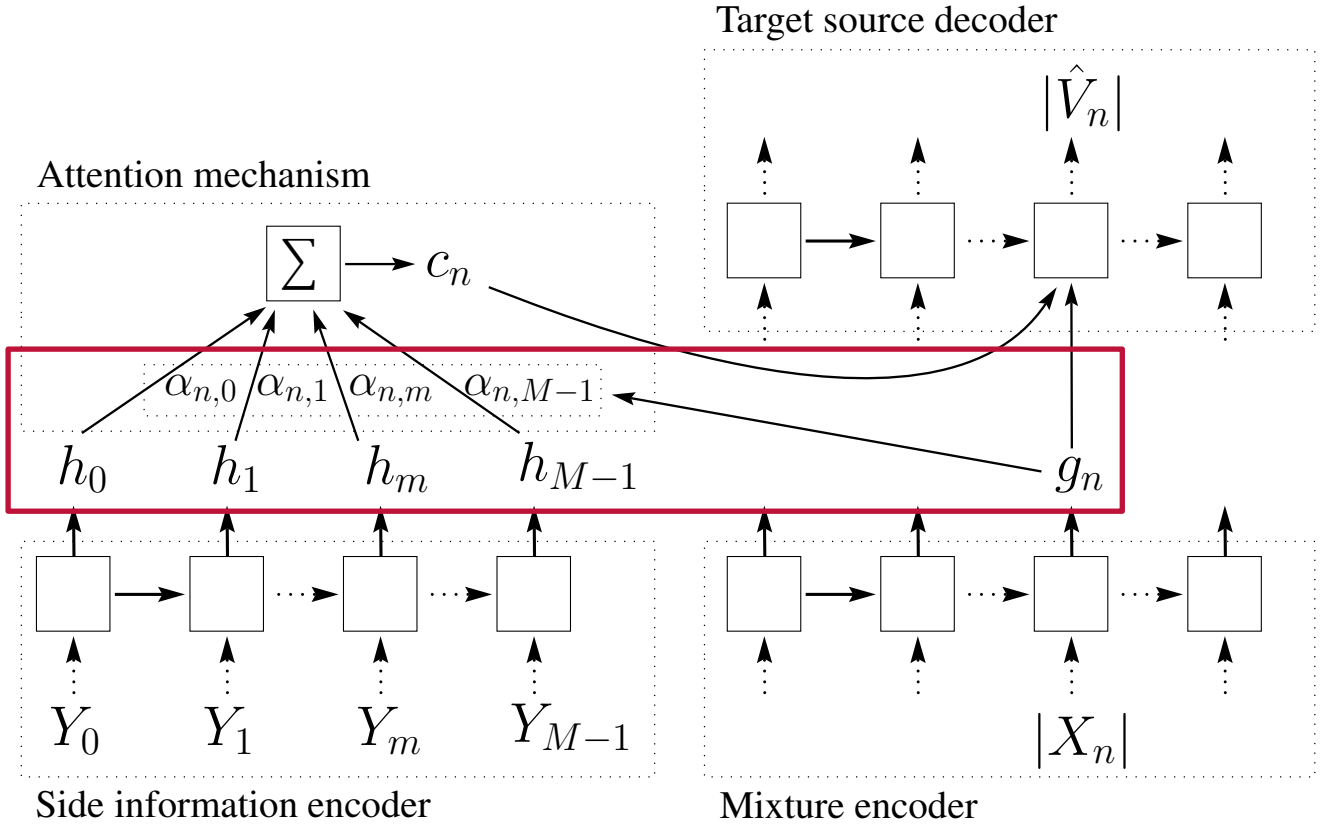
Proposed Model: Learn to Align and Separate Jointly



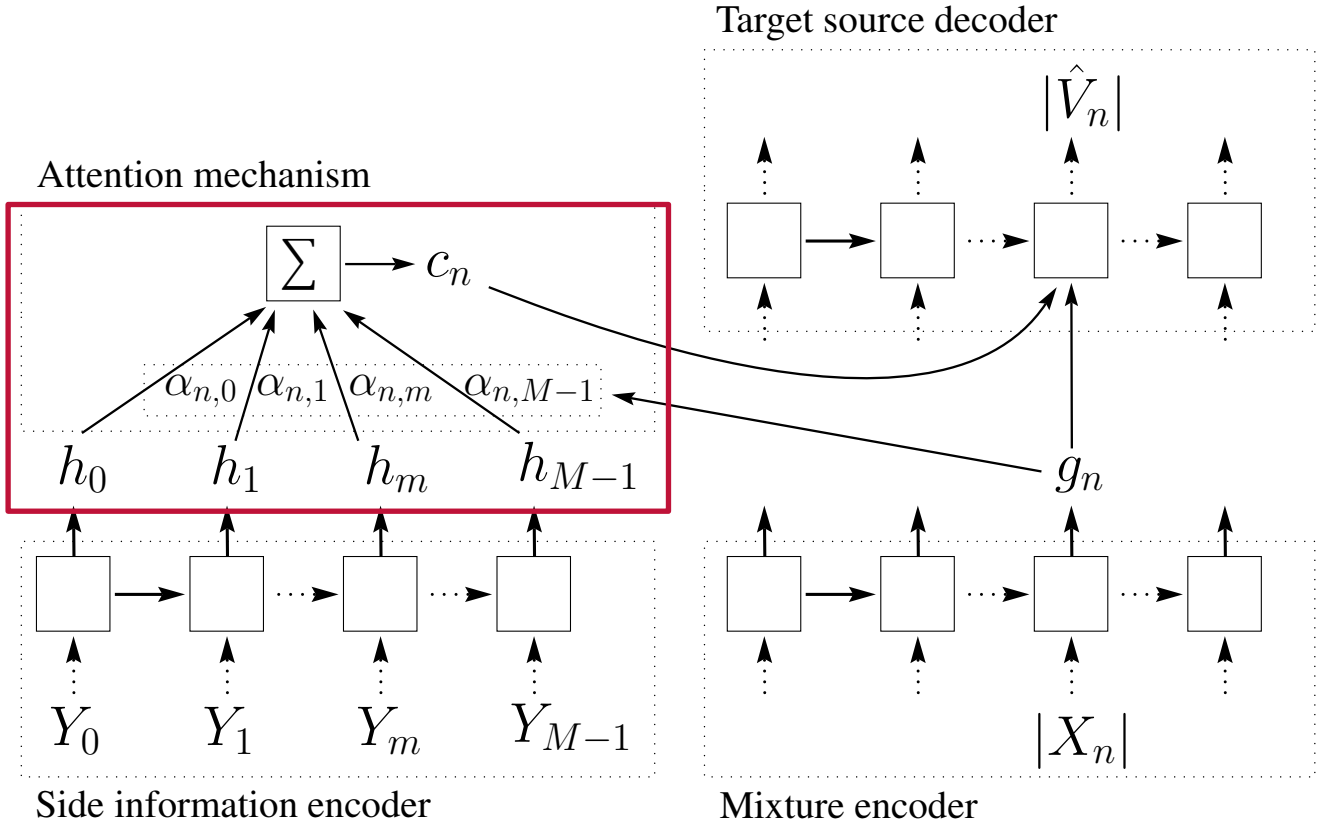
Proposed Model: Learn to Align and Separate Jointly



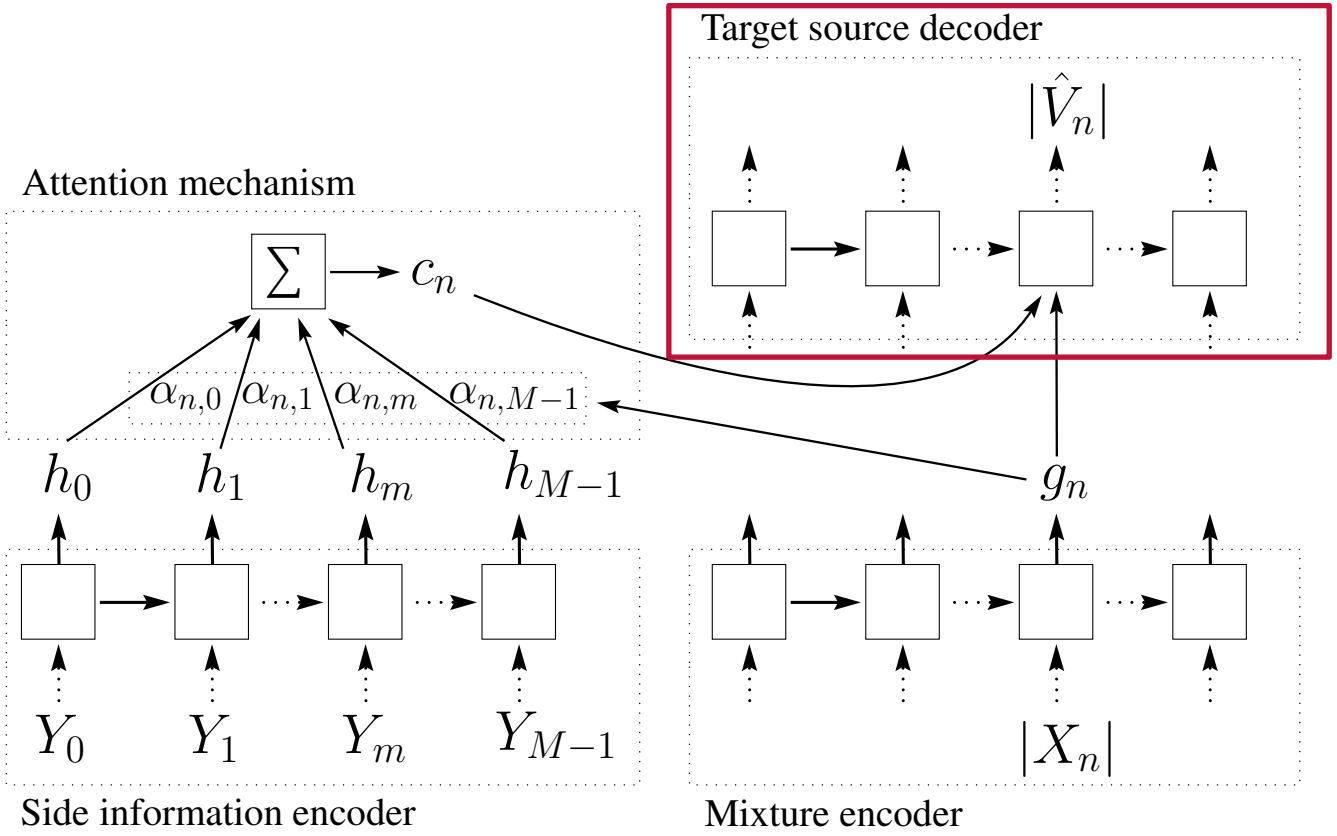
Proposed Model: Learn to Align and Separate Jointly



Proposed Model: Learn to Align and Separate Jointly



Proposed Model: Learn to Align and Separate Jointly



Audio Source Separation Quality Evaluation

- **SDR, SIR, SAR are not defined on evaluation frames with:**
  - **Silent target source**
  - **Silent prediction**

- **How much energy is in the prediction, when target is silent?**
  - **Predicted Energy at Silence (PES)**

$$\text{PES} := 10 \log_{10} \sum_{t=0}^{T-1} \hat{s}_j^2(t) \quad \text{if} \quad \sum_{t=0}^{T-1} s_j^2(t) = 0$$

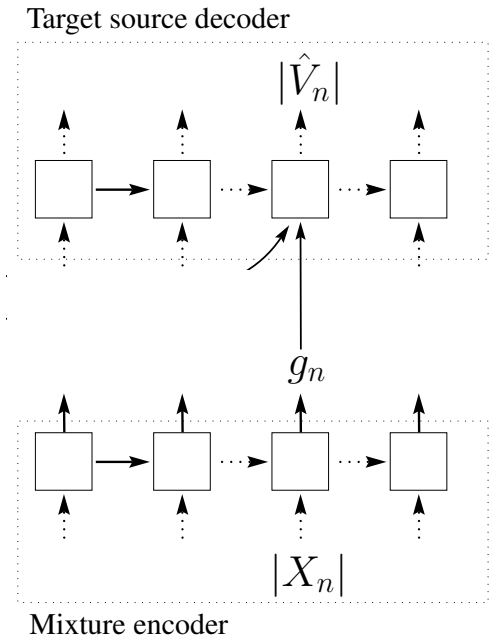
- **When the prediction is silent, how much energy is in the target?**
  - **Energy at Predicted Silence (EPS)**

$$\text{EPS} := 10 \log_{10} \sum_{t=0}^{T-1} s_j^2(t) \quad \text{if} \quad \sum_{t=0}^{T-1} \hat{s}_j^2(t) = 0$$

Singing Voice Separation with Artificial Side Information

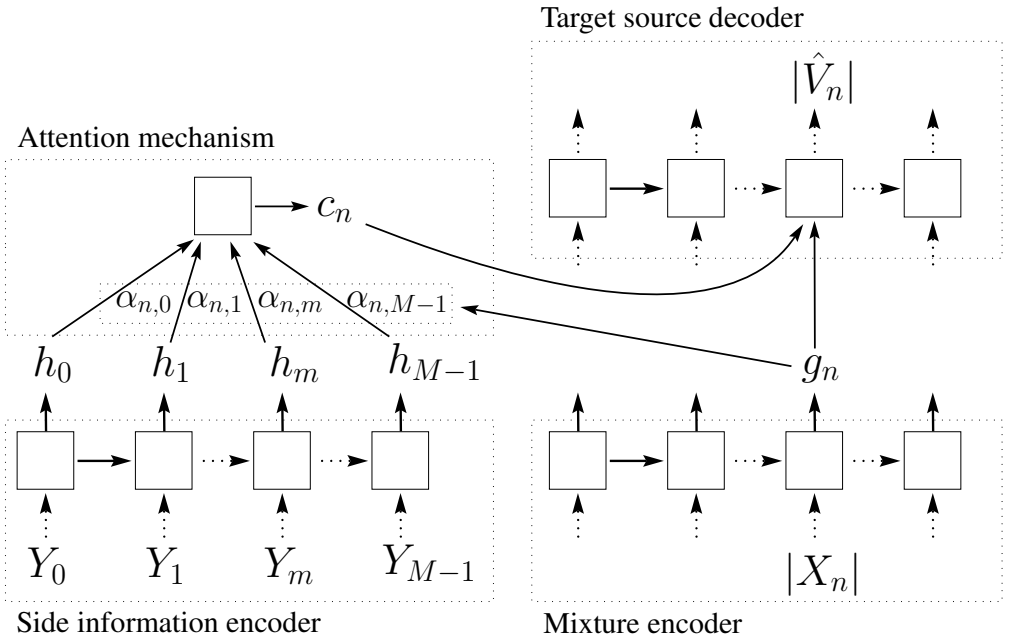
Singing Voice Separation with Artificial Side Information

- **Baselines:**
  - **BL1: No side information, no attention**



Singing Voice Separation with Artificial Side Information

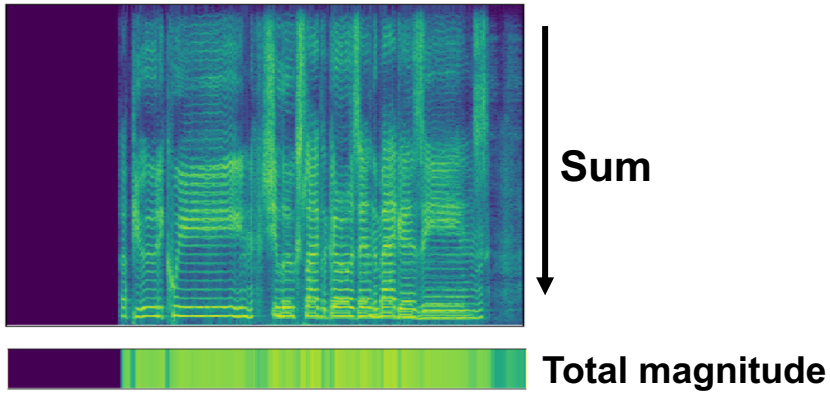
- **Baselines:**
  - **BL1: No side information, no attention**
  - **BL2: meaningless side information, complete model**



Singing Voice Separation with Artificial Side Information

- **Baselines:**
  - **BL1: No side information, no attention**
  - **BL2: meaningless side information, complete model**
- **Vocal activity information (A3.1)**

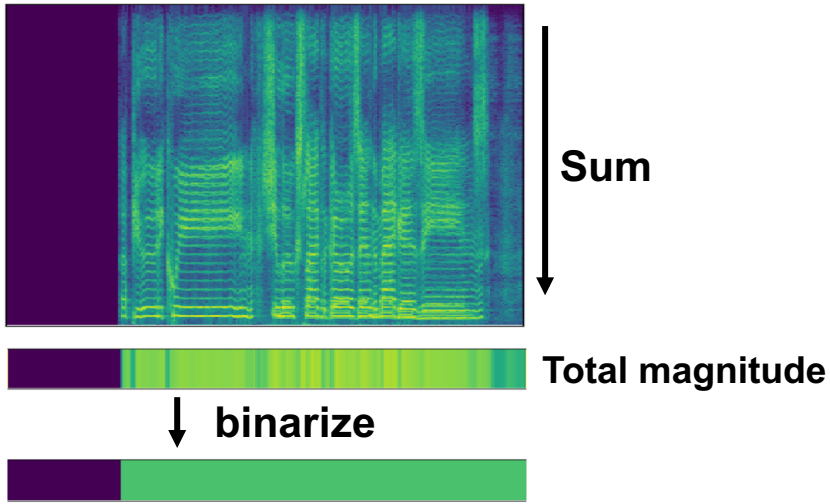
Ground truth vocals spectrogram



Singing Voice Separation with Artificial Side Information

- **Baselines:**
  - **BL1: No side information, no attention**
  - **BL2: meaningless side information, complete model**
- **Vocal activity information (A3.1)**
  - **Binary**

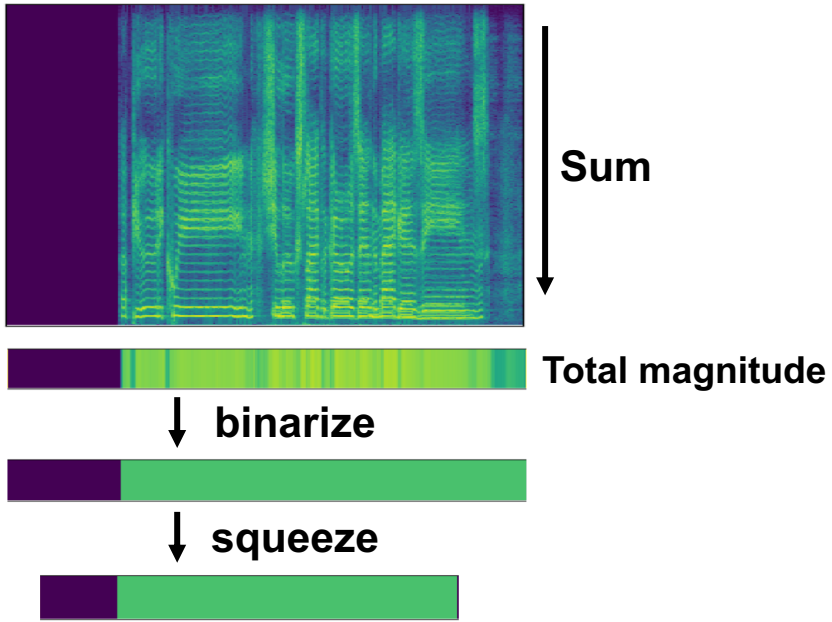
Ground truth vocals spectrogram



Singing Voice Separation with Artificial Side Information

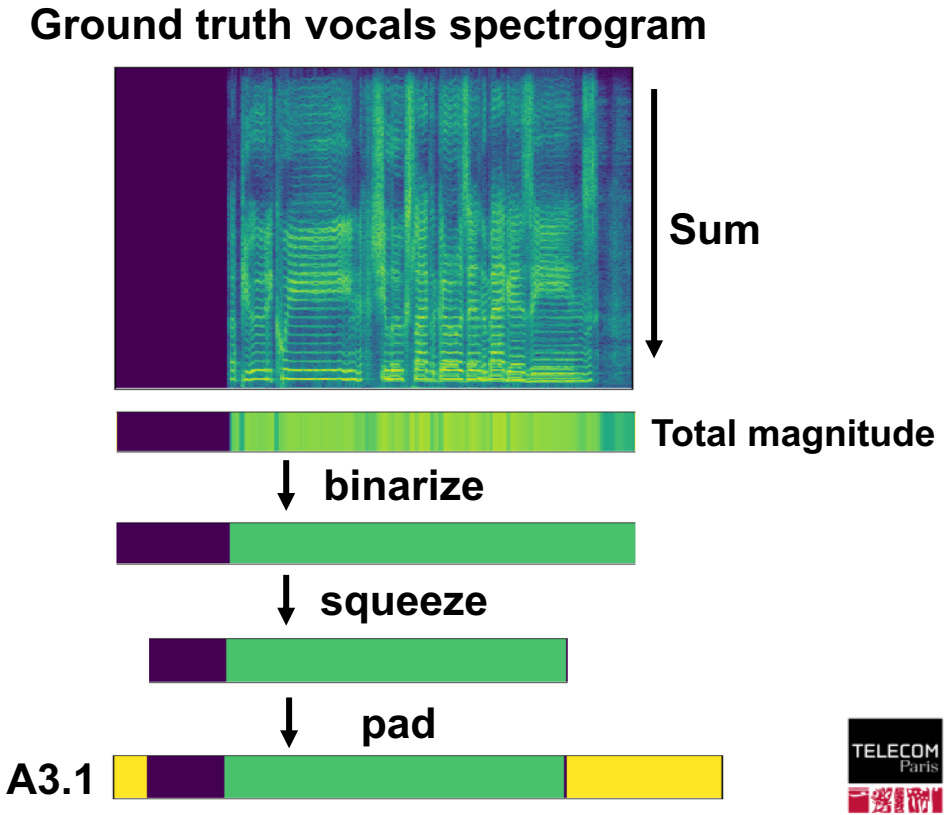
- **Baselines:**
  - **BL1: No side information, no attention**
  - **BL2: meaningless side information, complete model**
- **Vocal activity information (A3.1)**
  - **Binary**
  - **Different length than spectrogram**

Ground truth vocals spectrogram



Singing Voice Separation with Artificial Side Information

- **Baselines:**
  - BL1: No side information, no attention
  - BL2: meaningless side information, complete model
- **Vocal activity information (A3.1)**
  - Binary
  - Different length than spectrogram
  - Position of relevant information varies

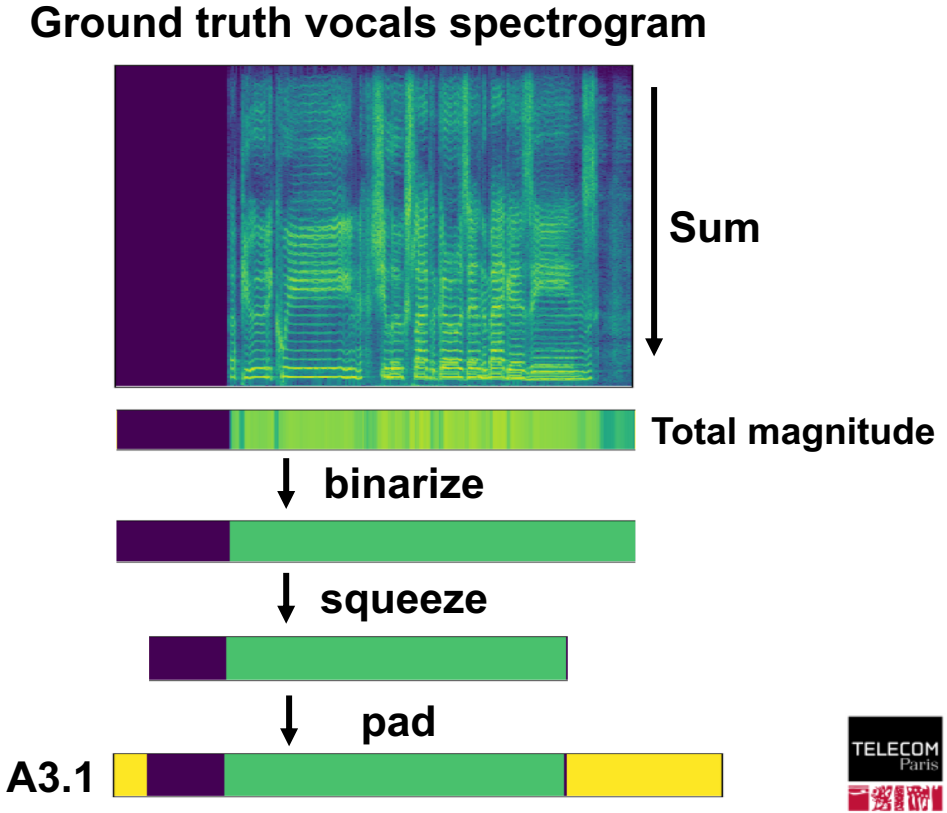


Singing Voice Separation with Artificial Side Information

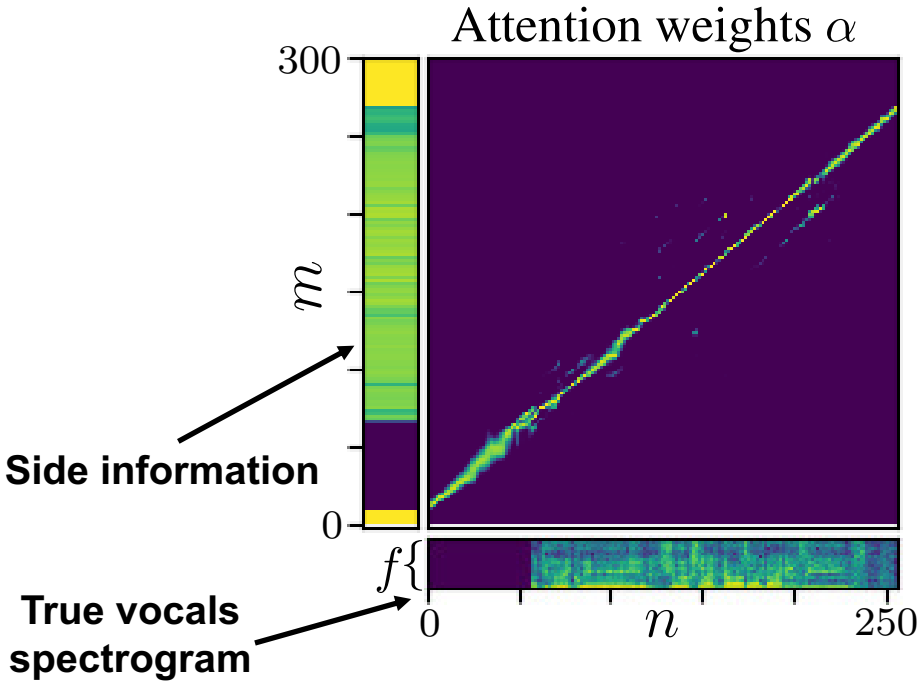
	SDR	SAR	SIR	PES	EPS
BL1	2.98	6.41	7.99	-44.89	-25.58
BL2	3.33	6.33	7.78	-43.78	-22.96
A3.1	3.16	6.17	7.75	**-85.20**	**-34.53**

Table 1: Evaluation results in dB for vocals on the MUSDB18 test set.

- **With voice activity information:**
  - **Improvement regarding silent parts**
  - **Standard metrics not improved**
- **The model could exploit this weak side information**

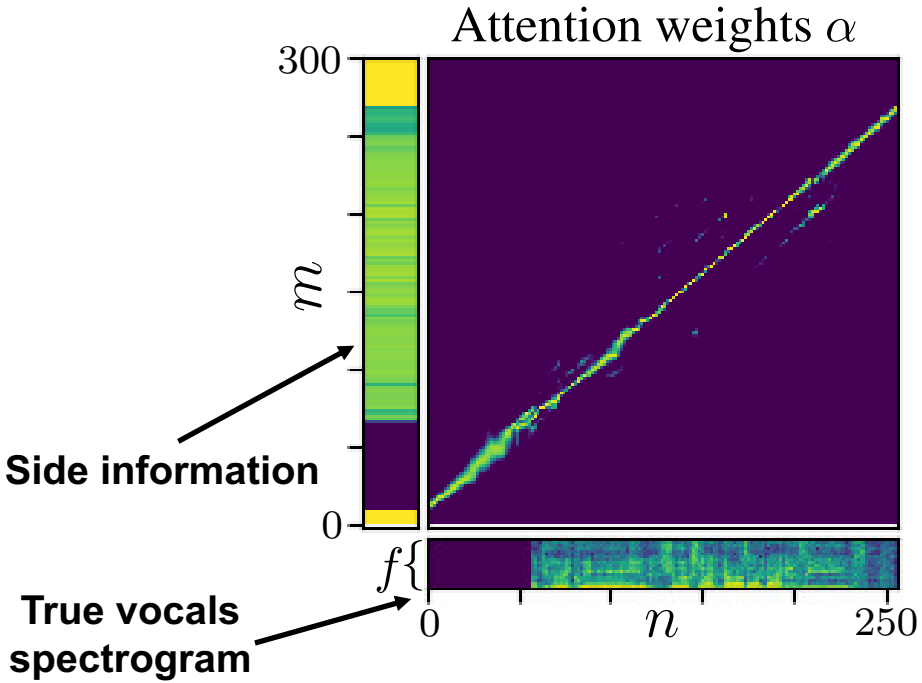


Learned Alignment by Training for Source Separation

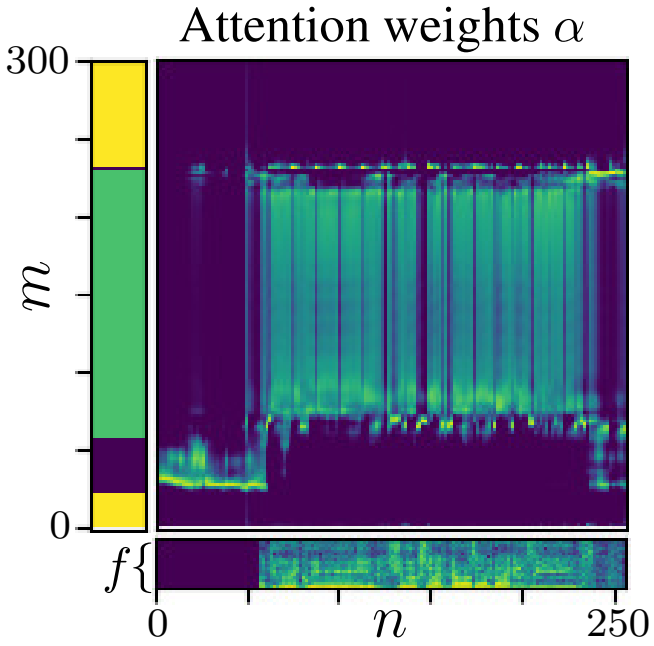


Vocal Magnitude Information

Learned Alignment by Training for Source Separation



Vocal Magnitude Information



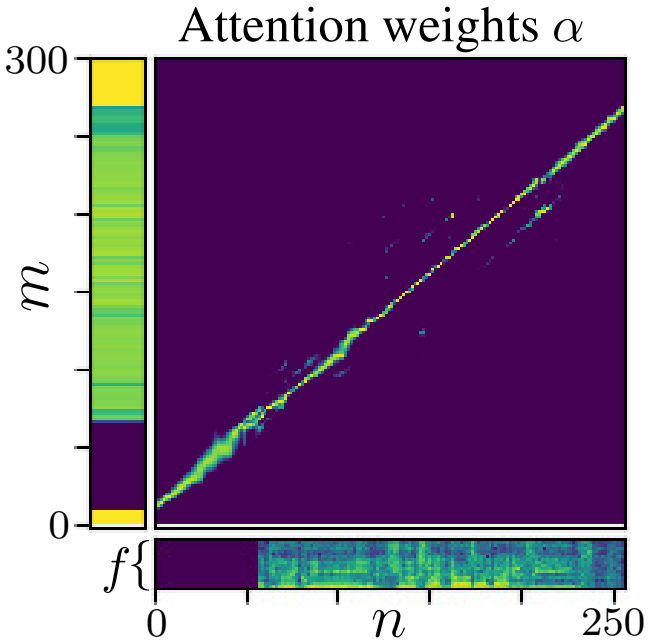
Binary Vocal Activity Information

Audio Examples

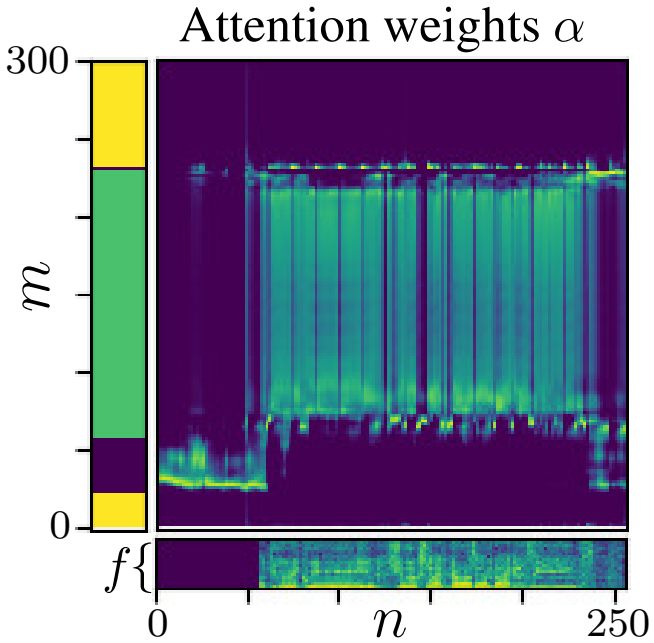
Mix

True Vocals

Baseline 1



Vocal Magnitude Information (M2)



Vocal Activity Information (A3.1)

Conclusion

- **Novel model for informed source separation**
- **Alignment of side information is learned by training for the separation task**
- **Also silent frames need to be evaluated for source separation**

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