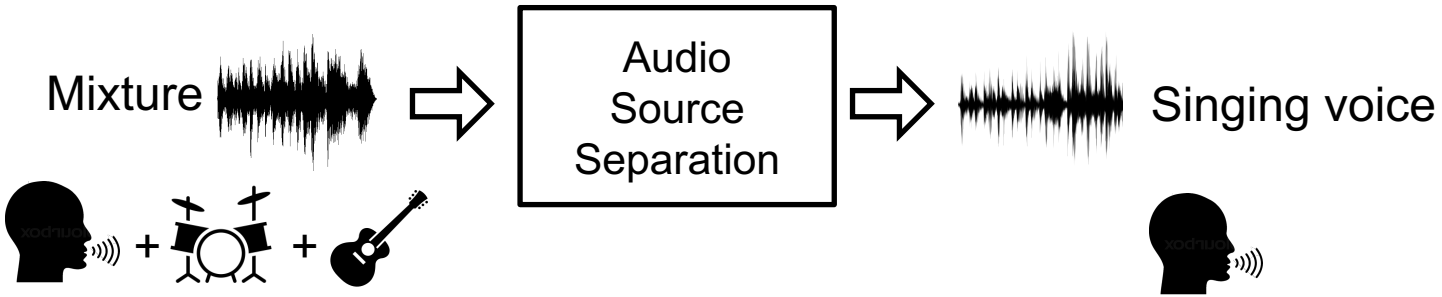
Text-Informed Singing Voice Separation and Phoneme Level Lyrics Alignment

Kilian Schulze-Forster[1]
Clement Doire,[2] Gaël Richard,[1] Roland Badeau[1]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris
[2] Audionamix

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.

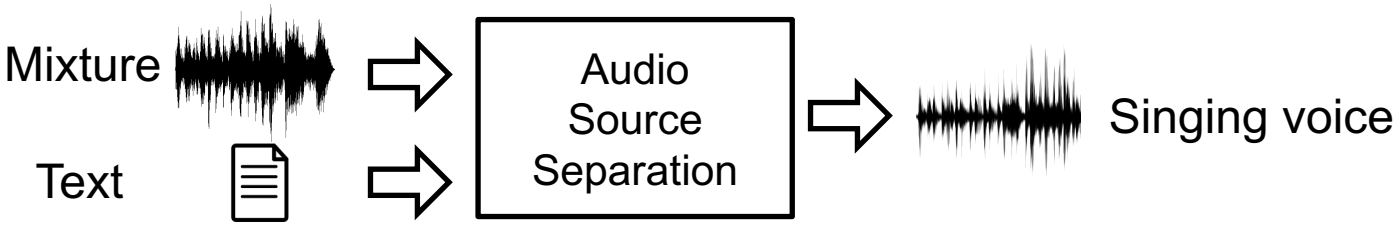
Introduction: Singing Voice Separation



- State-of-the-art: **Supervised deep learning** models[1], [2]
- Audio data for training are **difficult to obtain**
- Can singing voice separation be improved **without** access to **more audio data**?

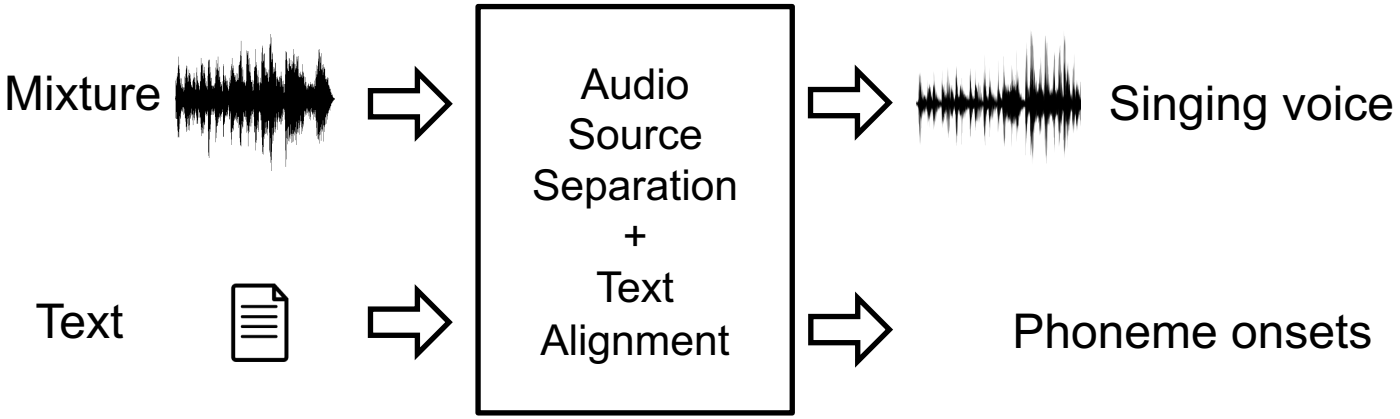
[1] Stöter, F. R., Uhlich, S., Liutkus, A., & Mitsufuji, Y. (2019). Open-Unmix - A Reference Implementation for Music Source Separation. *Journal of Open Source Software.*
[2] Défossez, A., Usunier, N., Bottou, L., & Bach, F. (2019). Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*.

Proposal: Text-Informed Singing Voice Separation



- **Challenge:**
  - Text and mixture signal must be aligned
  - Without singing voice separation as pre-processing

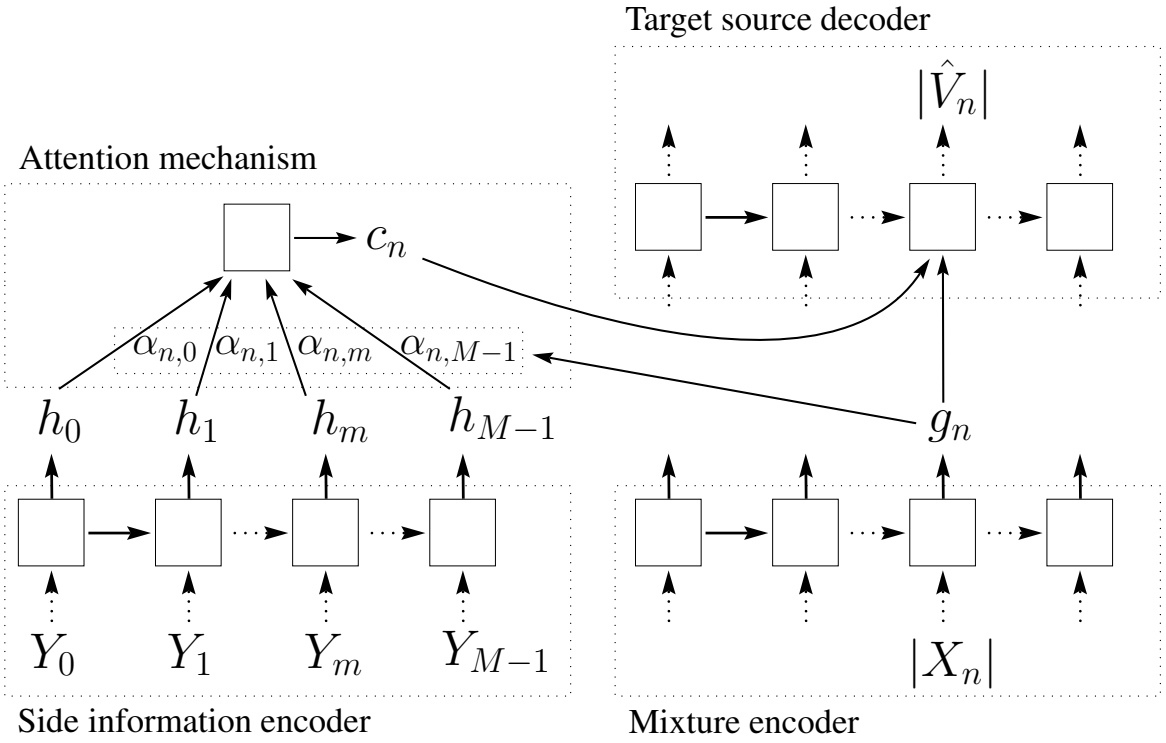
Text-Informed Singing Voice Separation and Joint Text Alignment



Schulze-Forster, K., Doire, C., Richard, G., & Badeau, R. (2019). Weakly informed audio source separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*

Schulze-Forster, K., Doire, C. S., Richard, G., & Badeau, R. (2020). Joint phoneme alignment and text-informed speech separation on highly corrupted speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

Proposed Model: Learn to Align and Separate Jointly

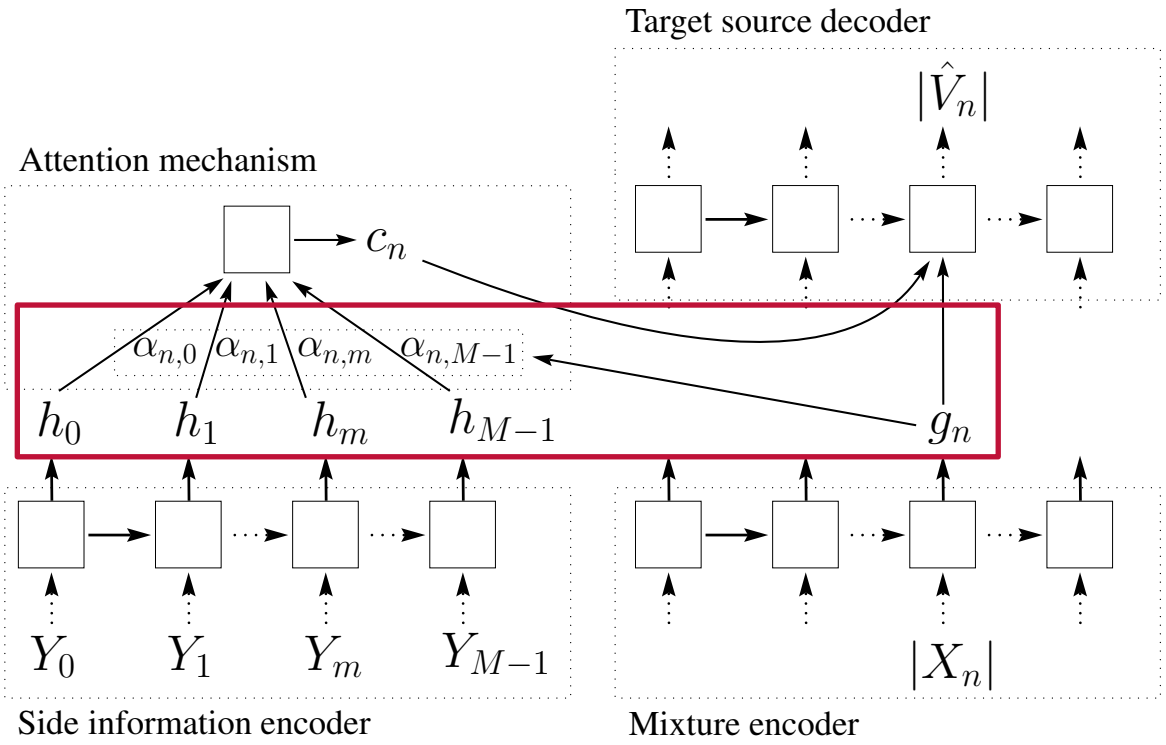


$$s_{n,m} = g_n^\top W h_m$$

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

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

Proposed Model: Learn to Align and Separate Jointly

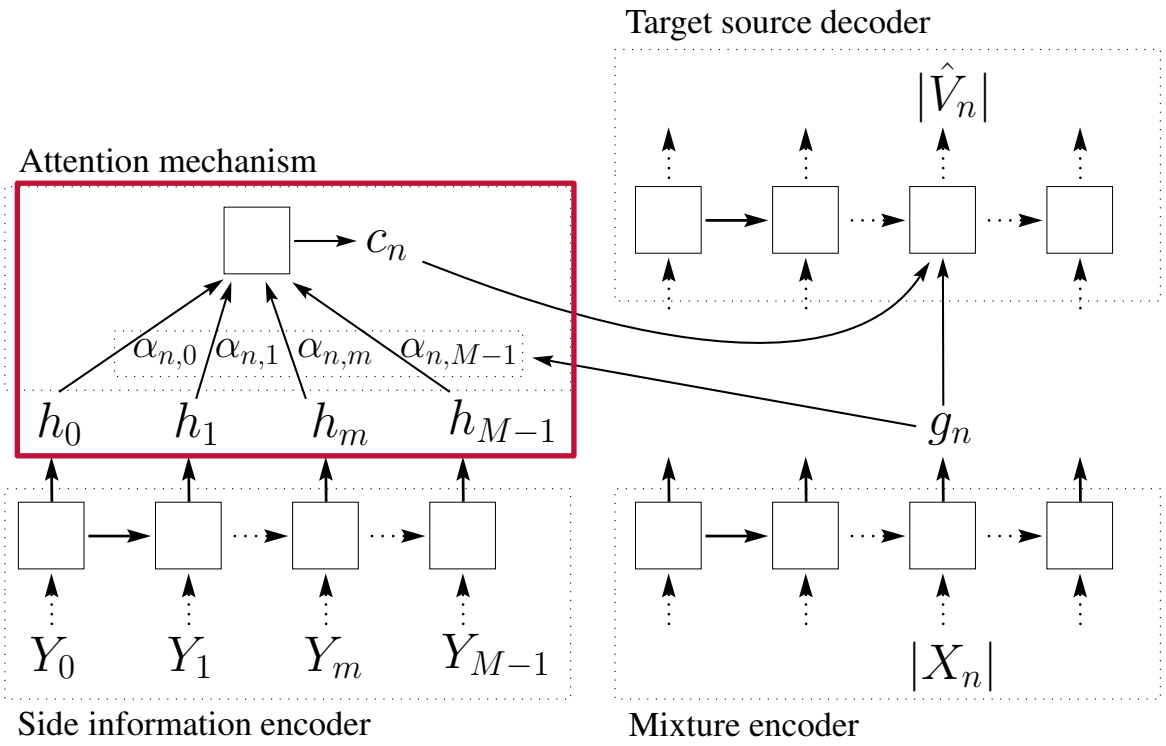


$$s_{n,m} = g_n^\top W h_m$$

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

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

Proposed Model: Learn to Align and Separate Jointly

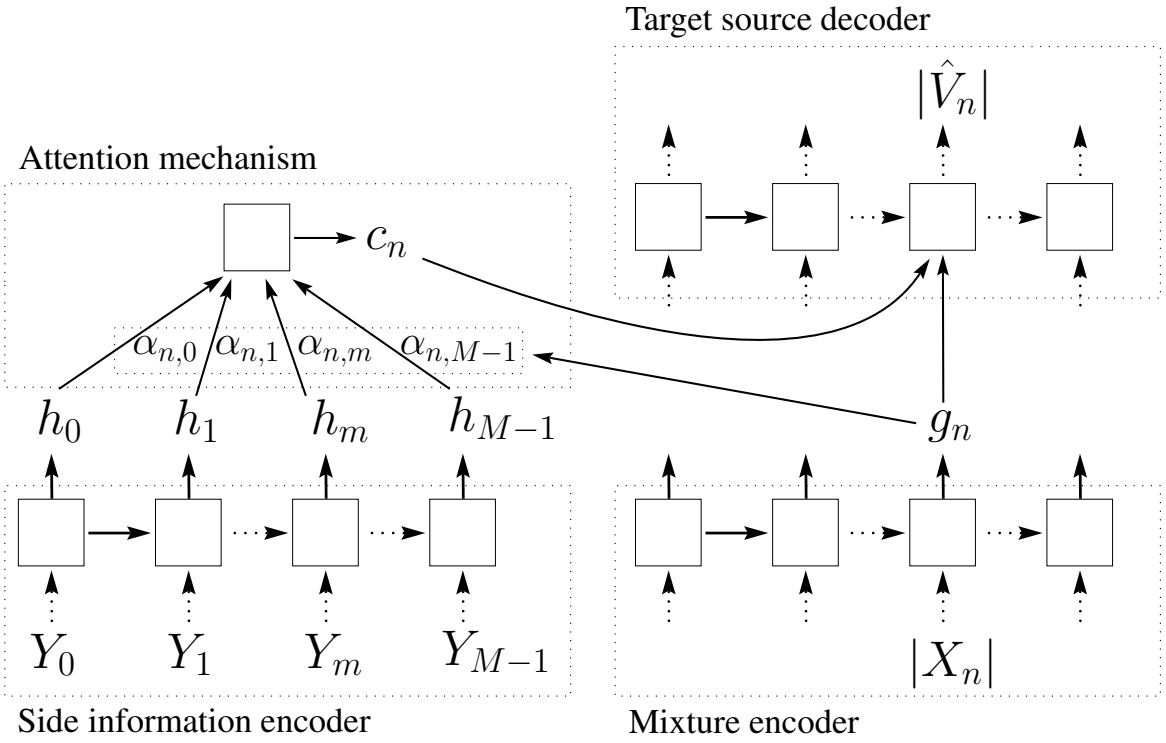


$$s_{n,m} = g_n^\top W h_m$$

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

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

Proposed Model: Learn to Align and Separate Jointly

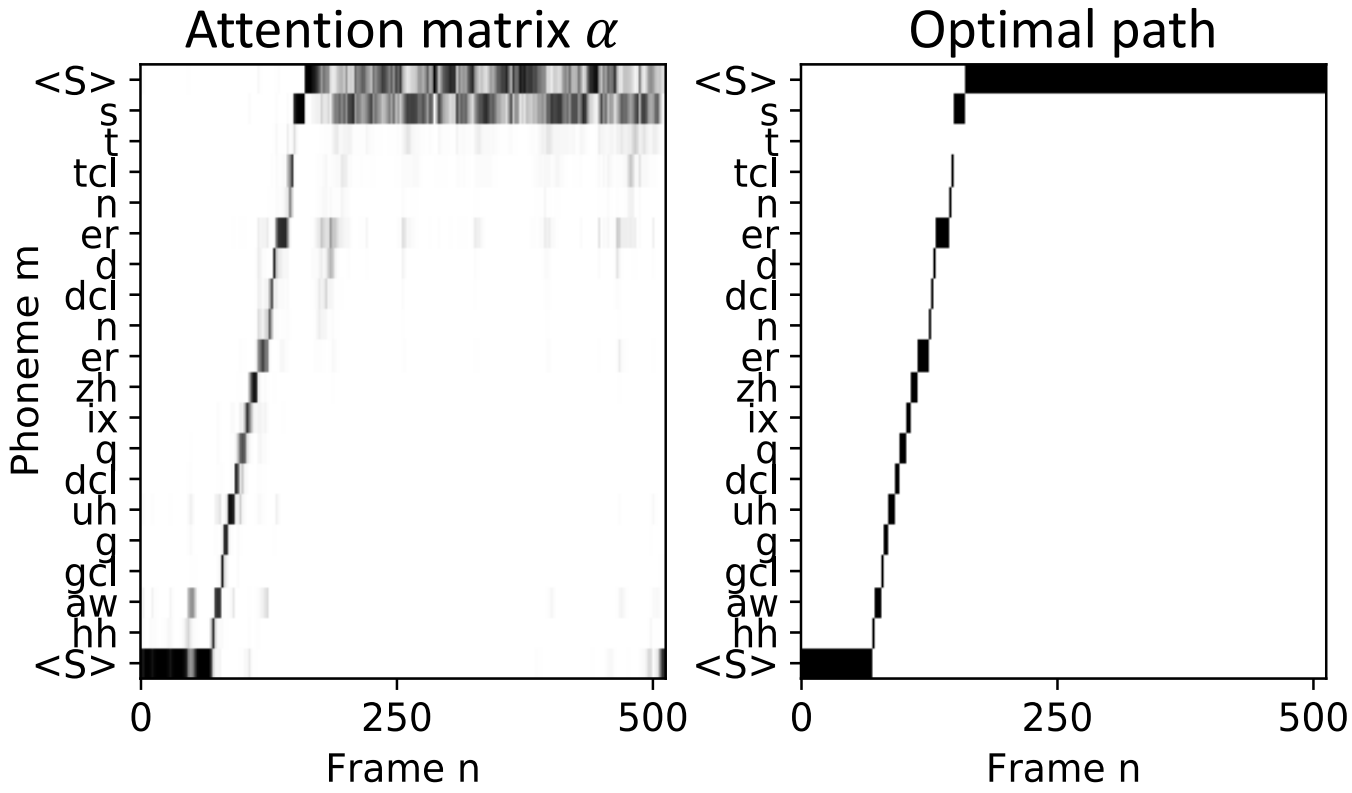


$$s_{n,m} = g_n^\top W h_m$$

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

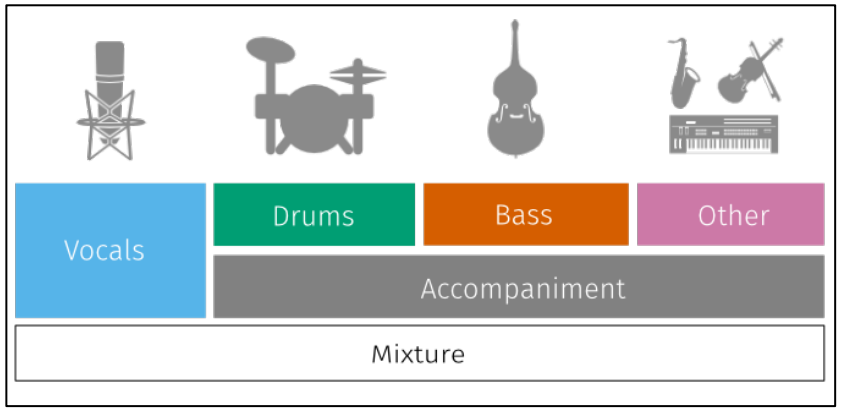
$$c_n = \sum_{m=0}^{M-1} h_m \alpha_{n,m}$$

Retrieving Phoneme Onsets from Attention Weights with DTW



Training Data: Mixtures, Clean Vocals, and Lyrics Transcripts

MUSDB18 [1] corpus



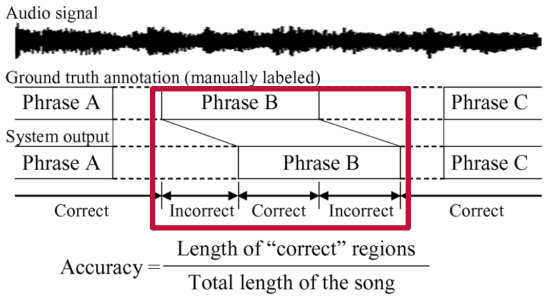
Coming soon: MUSDB lyrics extension

- Lyrics transcripts of the 141 songs in English
- Line level alignment
- Annotations for vocals track
  - 1 singer
  - 2+ singers, same text
  - 2+ singers, different text/ phonemes

[1] Rafii, Z., Liutkus, A., Stöter, F. R., Mimilakis, S. I., & Bittner, R. (2017). MUSDB18 - a corpus for music separation. (https://sigsep.github.io/datasets/musdb.html)

Results: Phoneme Level Lyrics Alignment

- **Test data**
  - NUS-48E corpus[1]
    - Solo singing recordings (1-3 minutes length)
    - **Accurate** phoneme transcripts with onsets
  - Mixed with MUSDB accompaniments
- **Baseline:** Montreal Forced Aligner [2]
- **Metric[3]:**



Method	PCAS [%]	SNR
ours	85.94	solo
baseline	77.94	singing
ours	84.66	5 dB
baseline	46.92	5 dB
ours	82.17	0 dB
baseline	25.61	0 dB
ours	76.21	-5 dB
baseline	10.03	-5 dB

PCAS = Percentage of Correcly Aligned Segments

[1] Duan, Zhiyan, et al. "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech." *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2013.

[2] McAuliffe, Michael, et al. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." *Interspeech*. 2017.

[3] Fujihara, Hiromasa, et al. "LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics." *IEEE Journal of Selected Topics in Signal Processing*, 2011.

Results: Text-Informed Singing Voice Separation

- **Test data:** MUSDB18 (only English songs)
- No improvement through text over baseline with **joint approach**
- **Improvements** through text in **sequential approach**:

Side Info	1 singer			2+ sing. 1 phon.			2+ sing. 2+ phon.		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
constant	4.77	9.51	7.15	4.93	9.38	6.85	4.19	9.05	5.82
voice activity	4.74	9.17	6.83	4.55	9.14	6.94	3.75	8.62	5.70
phonemes	5.08	10.41	6.82	4.89	10.21	6.71	3.85	9.82	5.03

Evaluation scores in dB. Median over evaluation frames.

Listening Examples

Conclusion

- Model for joint text alignment and text-informed voice separation
  - **Attention mechanism** between two encoders for **unsupervised alignment**
  - Separation facilitates **alignment with mixtures**
  - Sequential approach enables **improving singing voice separation** through text information
- MUSDB extension with **lyrics and vocals annotations**

kilian.schulze-forster@telecom-paris.fr