Deep neural Network for audio and music transformations

Journée Statistique & Informatique pour la Science des Données à Paris-Saclay

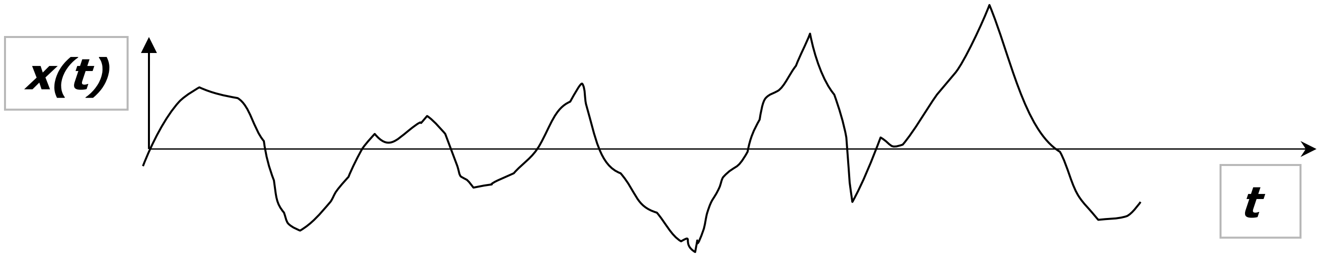
Gaël RICHARD
Télécom Paris, Institut polytechnique de Paris
February 5, 2021

Content

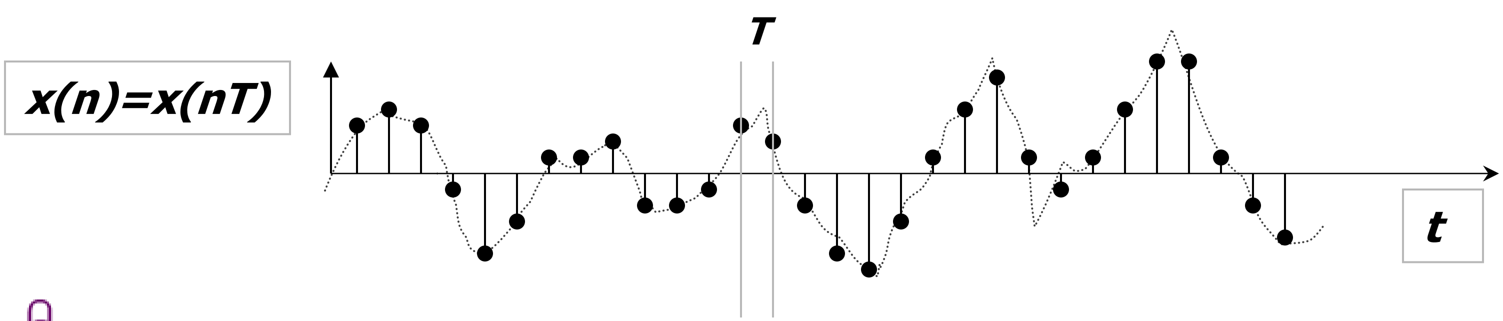
- **Introduction:**
  - The audio signal and its representations (e.g. spectrogram)
- **Deep learning for audio**
  - Differences with Images
  - Specific architectures for Audio
- **A focus on two application examples**
  - Text-Informed singing voice separation
  - Music Style transfer

The audio signal …

- Let x(t) be a continuous signal (e.g. captured by a microphone):



- Let x(nT) be the discrete signal sampled at time *t=nT*



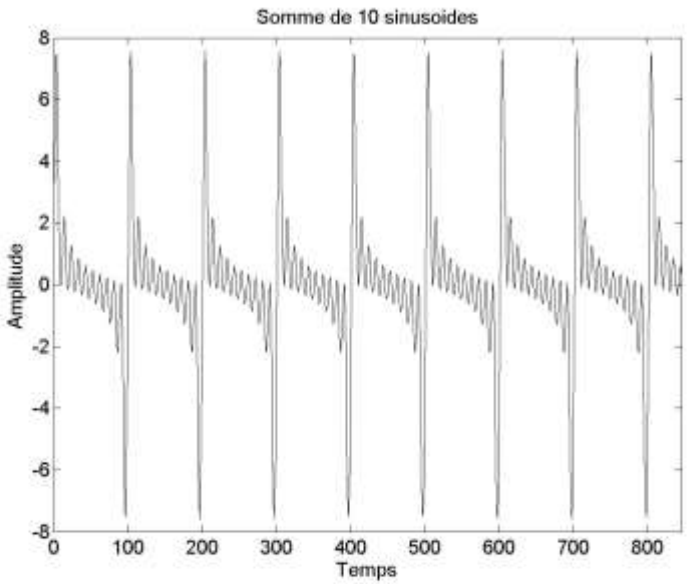
Time-Frequency representation

- **Fourier Transform**

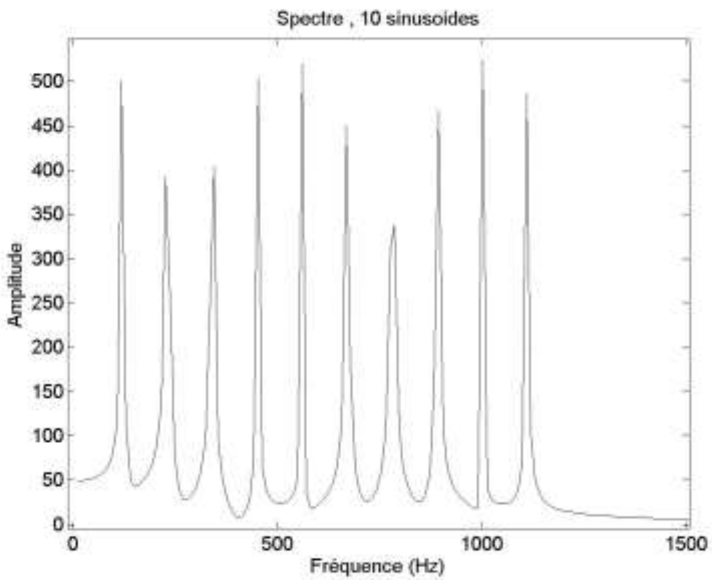
$$X_k = \sum_{n=0}^{N-1} x_n e^{-2j\pi nk/N}$$

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{2j\pi nk/N}$$

x_n

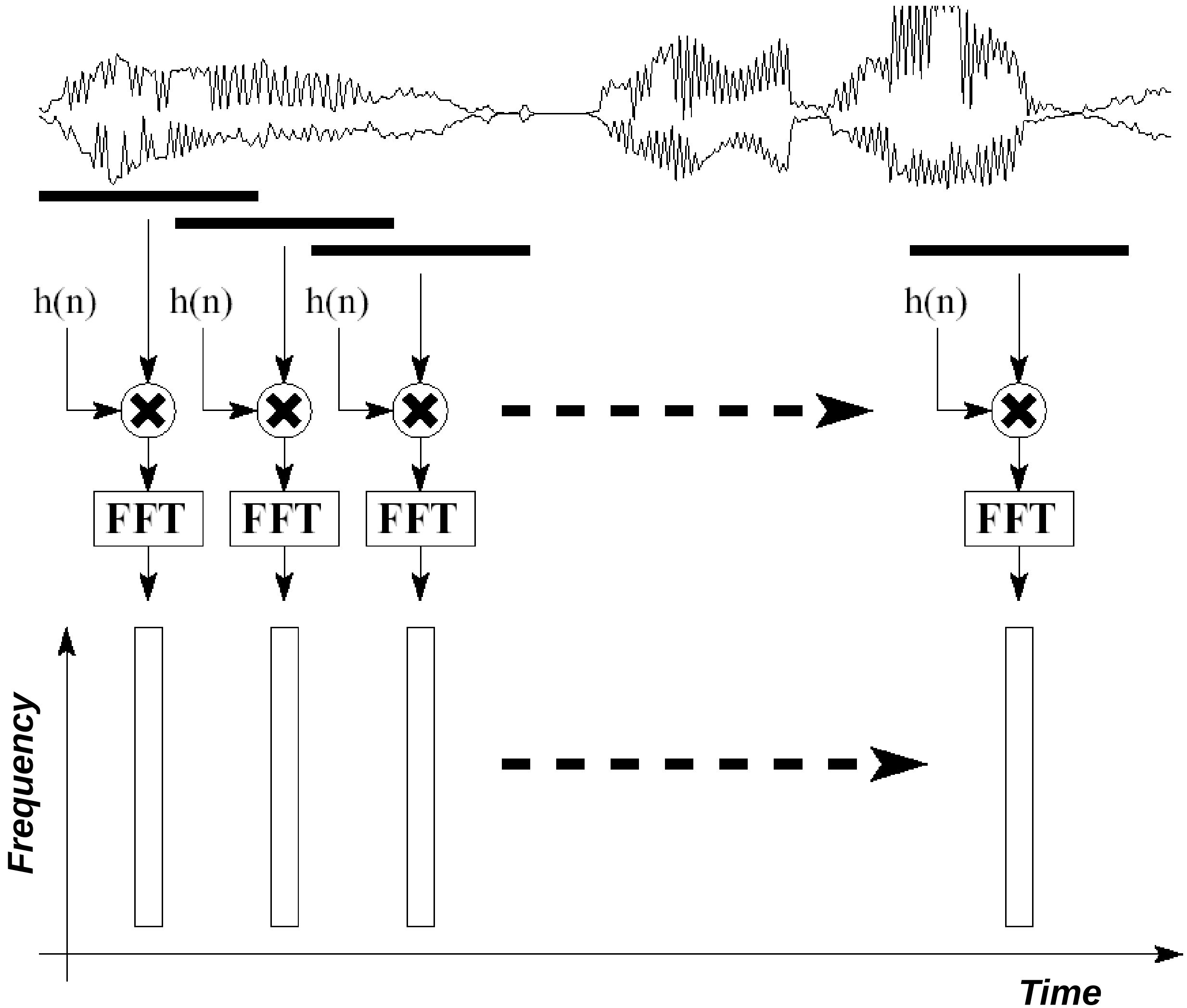


$|X_k|$



Spectral analysis of an audio signal (1)

(drawing from J. Laroche)

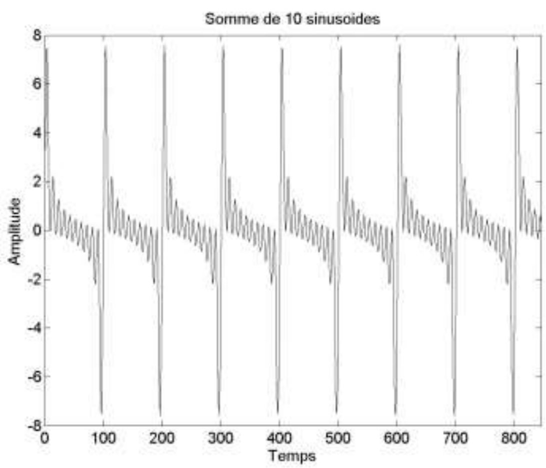


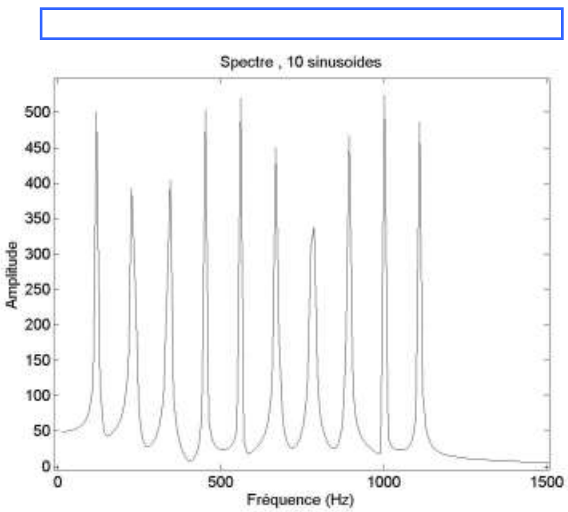
Spectral analysis of an audio signal (2)

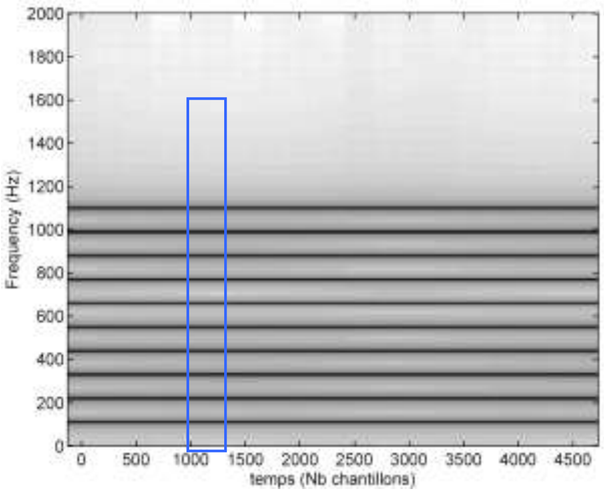
Spectrogram

x_n

$|X_k|$





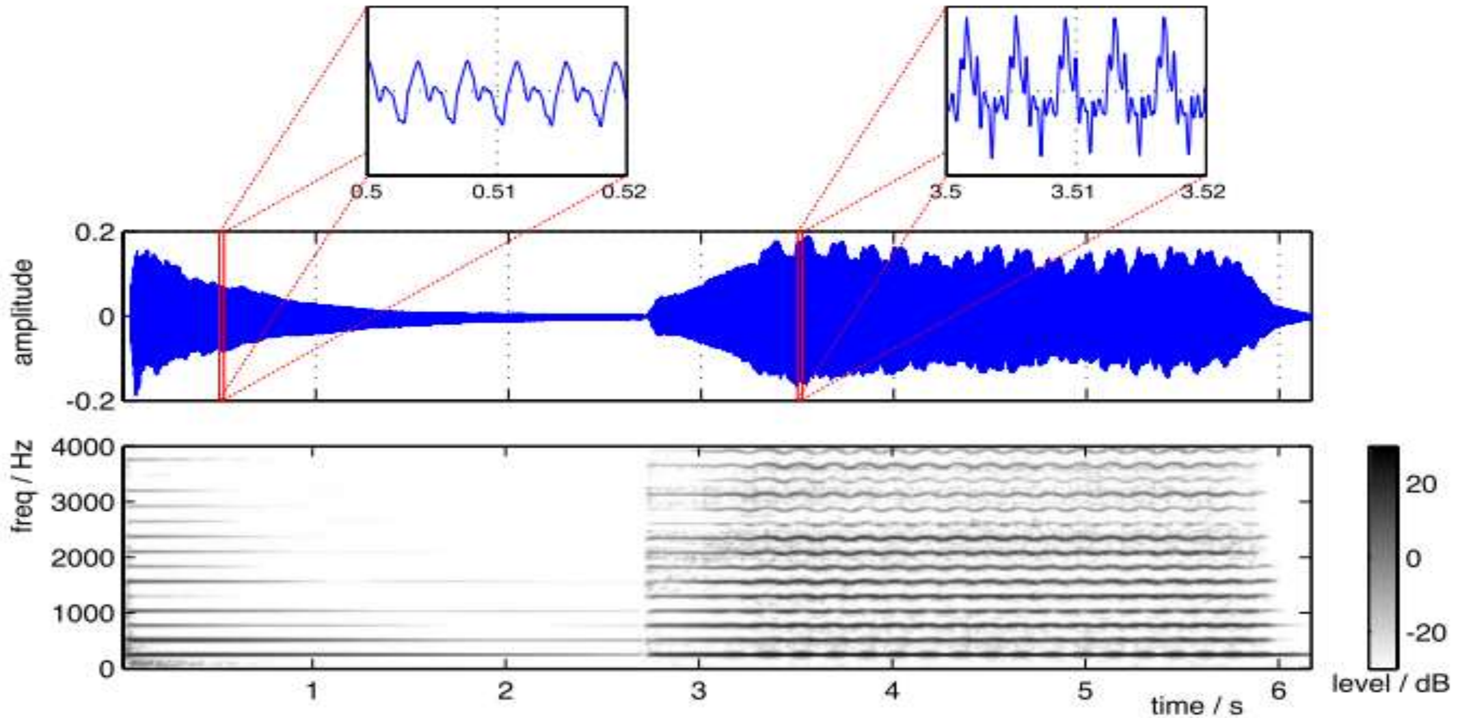


Audio signal representations

- **Example on a music signal: note C (262 Hz) produced by a piano and a violin.**

Temporal Signal

Spectrogram

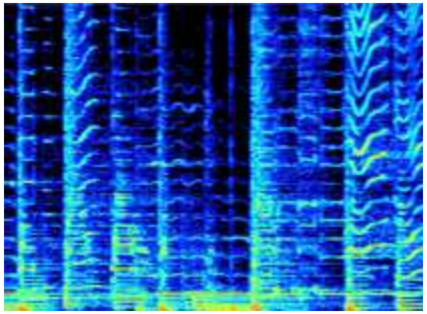


From M. Mueller & al. « Signal Processing for Music Analysis, IEEE Trans. On Selected topics of Signal Processing, oct. 2011

Deep learning for audio

Differences between an image and audio representation

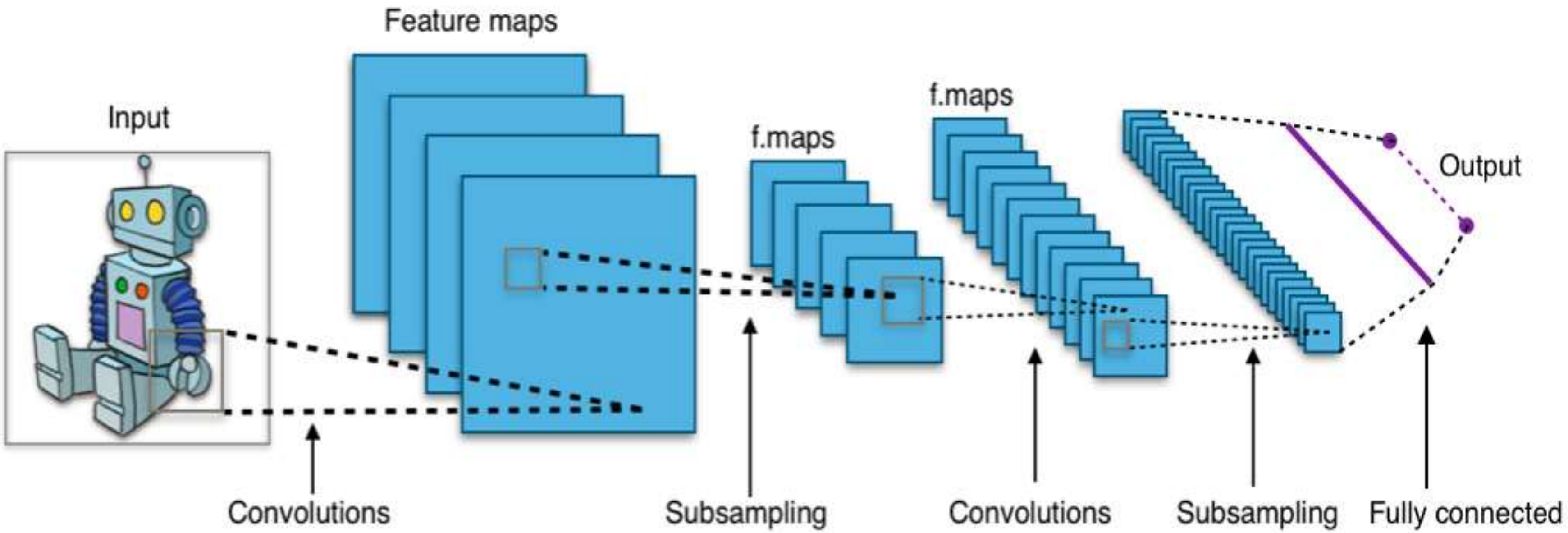
- x and y axes: **same concept** (spatial position).
- Image elements (cat's ear) : **same meaning** independently of their positions over x and y.
- **Neighbouring pixels** : often correlated, often belong to the same object
- **CNN are appropriate :**
  - Hidden neurons locally connected to the input image,
  - Shared parameters between various hidden neurons of a same feature map
  - Max pooling allows spatial invariance



- x and y axes: **different concepts** (time and frequency).
- Spectrogram elements (e.g. a time-frequency area representing a sound source): **same meaning** independently in time **but not over frequency**.
- No invariance over y (even with log-frequency representations): neighboring pixels of a spectrogram are not necessarily correlated since an harmonic sound can be distributed overt he whole frequency in a sparse way
- **CNN not as appropriate than it is for natural images**

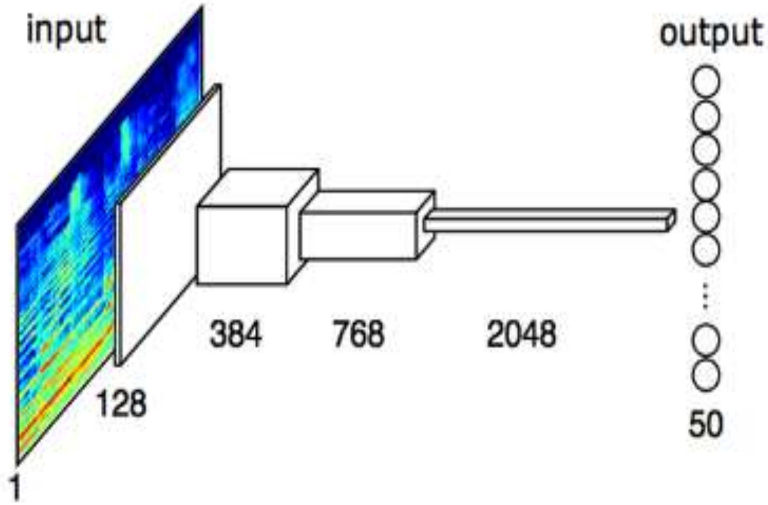
G. Peeters, G. Richard, « Deep learning for audio» , Multi-faceted Deep Learning: Models and Data, Edited by Jenny Benois-Pineau, Akka Zemmari, Springer-Verlag, 2021 (to appear)

A typical CNN



From https://en.wikipedia.org/wiki/Convolutional_neural_network

Music automatic tagging with CNN



Tags are include:
- **emotion** (sad, anger, happy),
- **genre** (jazz, classical)
- **instrumentation** (guitar, strings, vocal, instrumental).

FCN-4
Mel-spectrogram *(input: 96×1366×1)*
Conv 3×3×128
MP (2, 4) *(output: 48×341×128)*
Conv 3×3×384
MP (4, 5) *(output: 24×85×384)*
Conv 3×3×768
MP (3, 8) *(output: 12×21×768)*
Conv 3×3×2048
MP (4, 8) *(output: 1×1×2048)*
Output 50×1 (sigmoid)

- Good results,…. despite the pure « image based » architecture (due to mel-spectrogram ?)
- **But can be improved…..**

From: K. Choi & al. Automatic tagging usingdeep convolutional neural networks. InProc. of ISMIR (International Societyfor Music Information Retrieval), New York, USA, 2016.

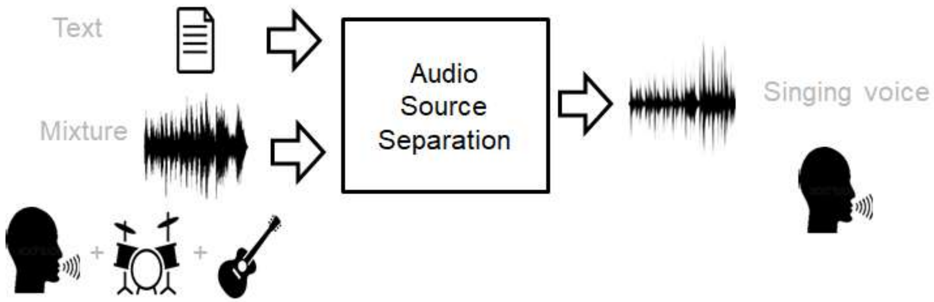
Deep learning for audio signals

- **Some interesting or popular directions**
  - Use « musically motivated » CNN
    - « Horizontal filters » (or temporal) or « vertical filters » (frequency)
  - Use different input representations
    - Mel-spectrograms, Constant-Q transform (CQT),
    - Non-negative Matrix factorisation (NMF), waveform,
  - To represent the sequential aspect of the audio signal
    - Use of Temporal NN, Recurrent NN
    - Exploit specific units to face the vanishing gradient problem
      - Long-Short term Memory (LSTM), Gated Recurrent Units (GRU),..
  - To use generative models (GANs,…)
  - To use Attention mechanisms

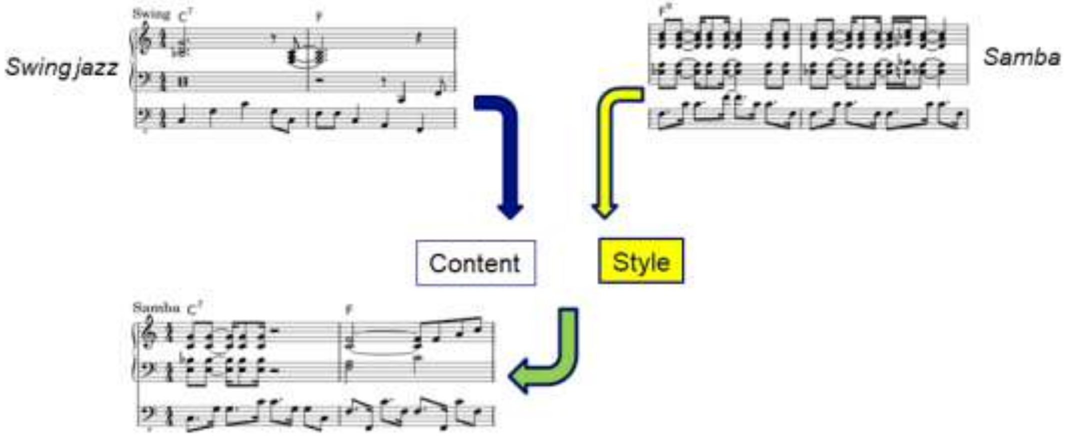
J.Pons & al., Experimenting with musically motivated convolutional neural networks. InProc. of IEEE CBMI, 2016

Illustration with two applications

- **Text-Informed singing voice (or speech) separation**



- **Music style transfer**



Text-informed singing voice (or speech) separation

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

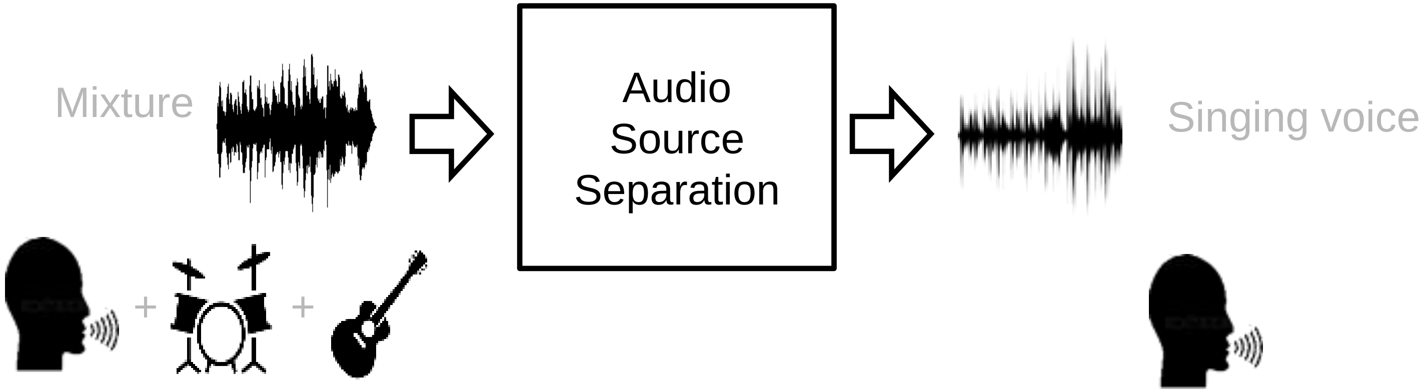
MIPFrontiers

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.

TELECOM Paris

IP PARIS

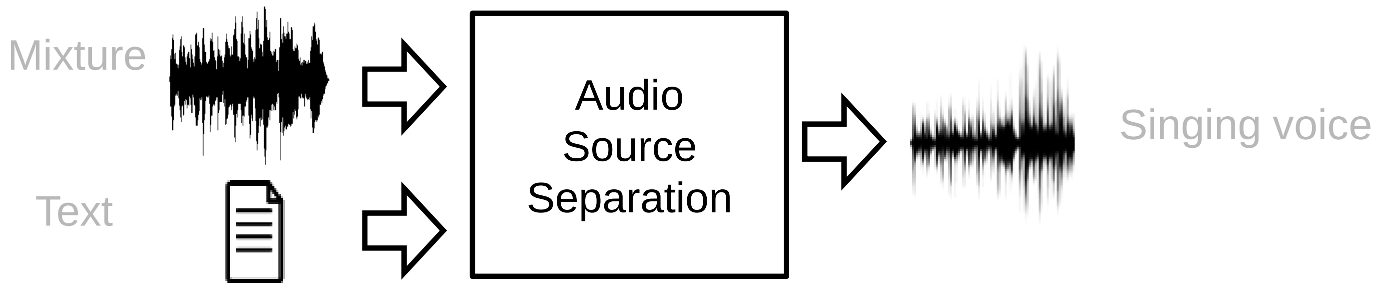
Introduction: Singing Voice Separation



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

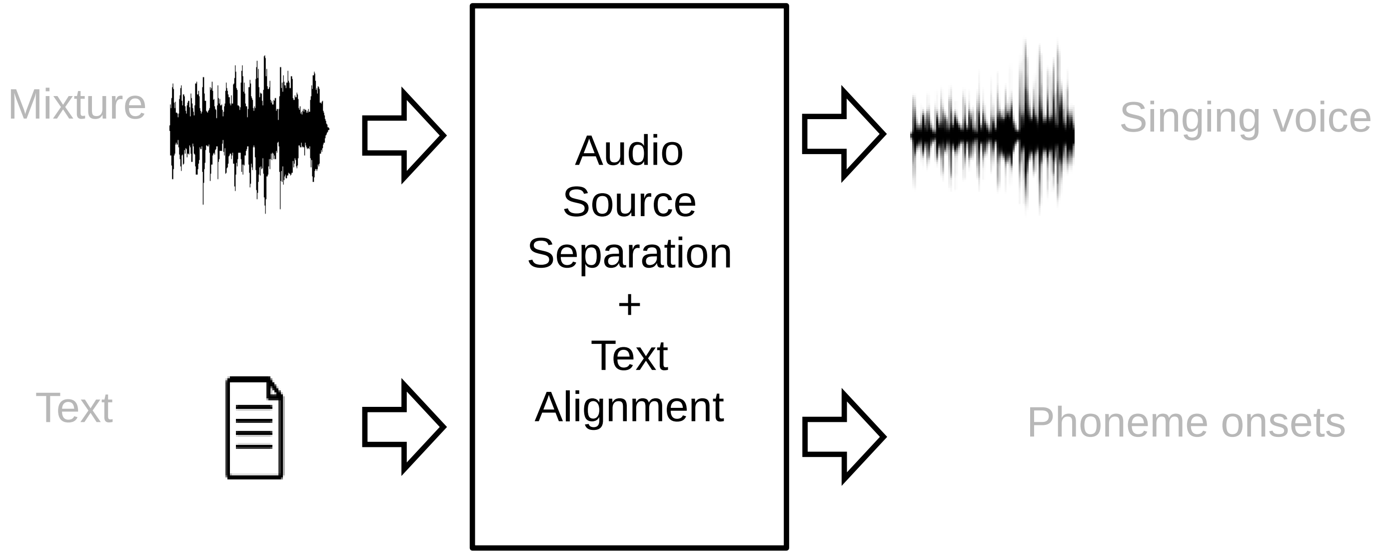
Stöter, F. R., Uhlich, S., Liutkus, A., & Mitsufuji, Y. (2019). Open-Unmix - A Reference Implementation for Music Source Separation. *Journal of Open Source Software.*
Défossez, A., Usunier, N., Bottou, L., & Bach, F. (2019). Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed.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

Proposal: Text-Informed Singing Voice Separation



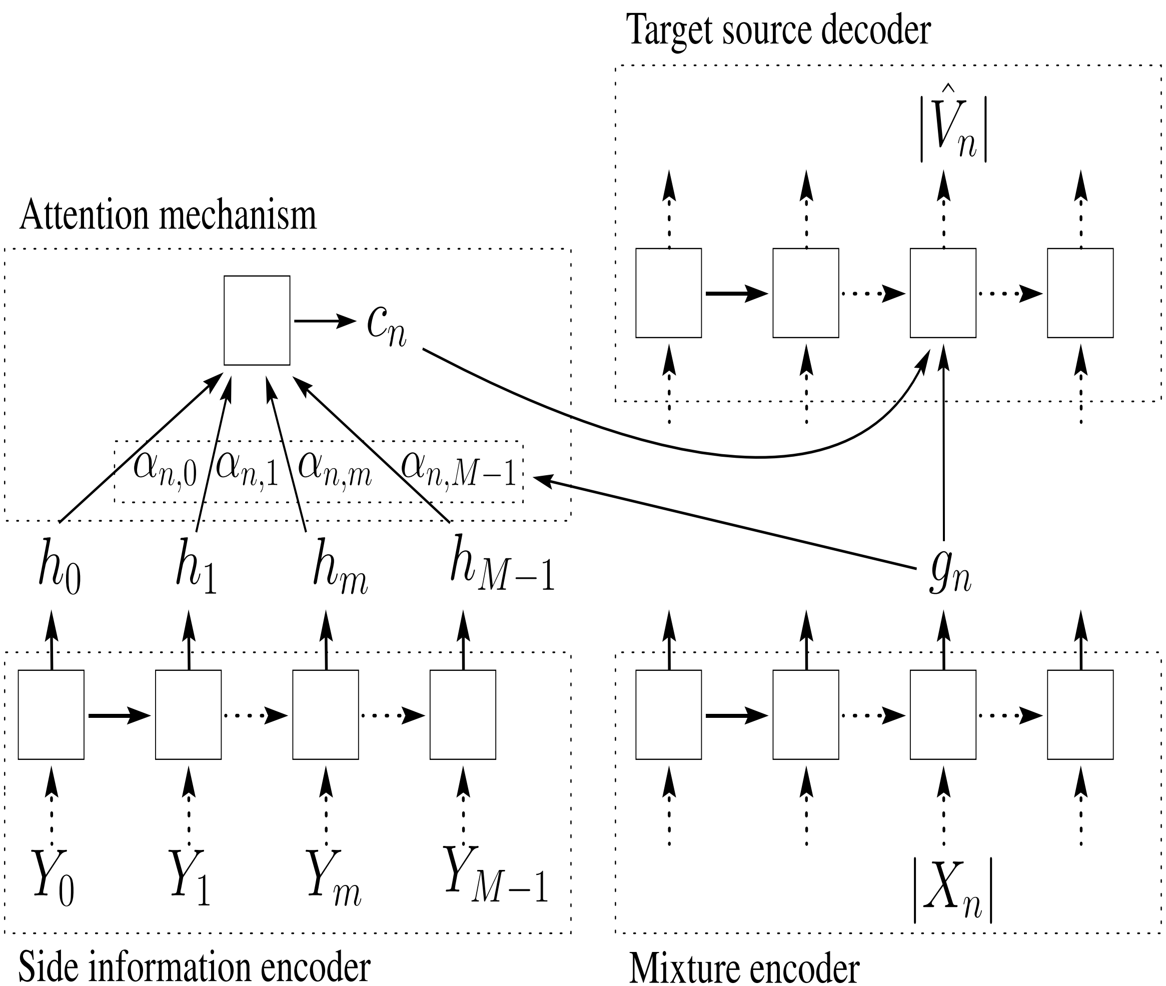
- **Challenge:**
  - Text and mixture signal must be aligned
  - Without singing voice separation as pre-processing

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

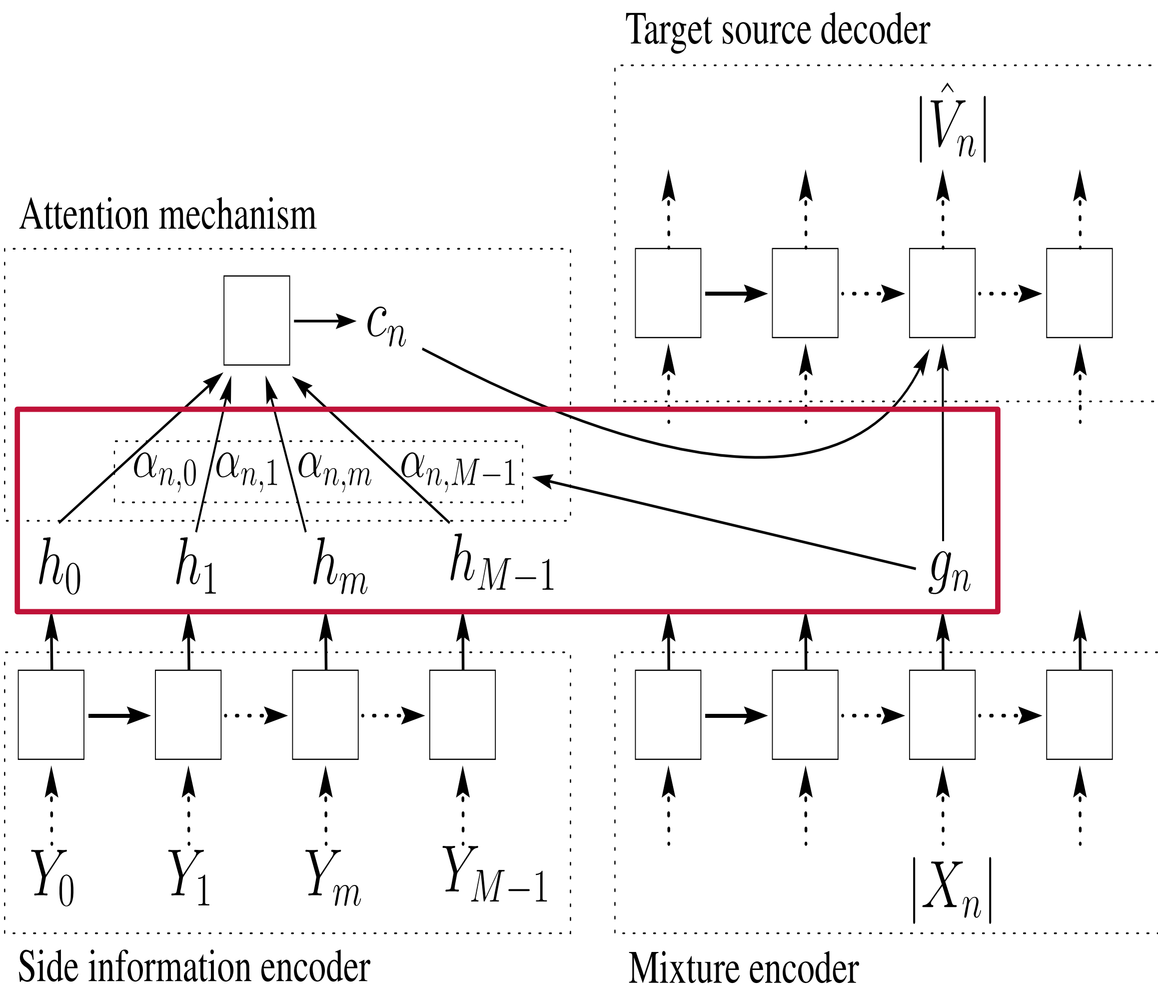


$$s_{n,m} = g_n^\top W h_m$$

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

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

Proposed Model: Learn to Align and Separate Jointly

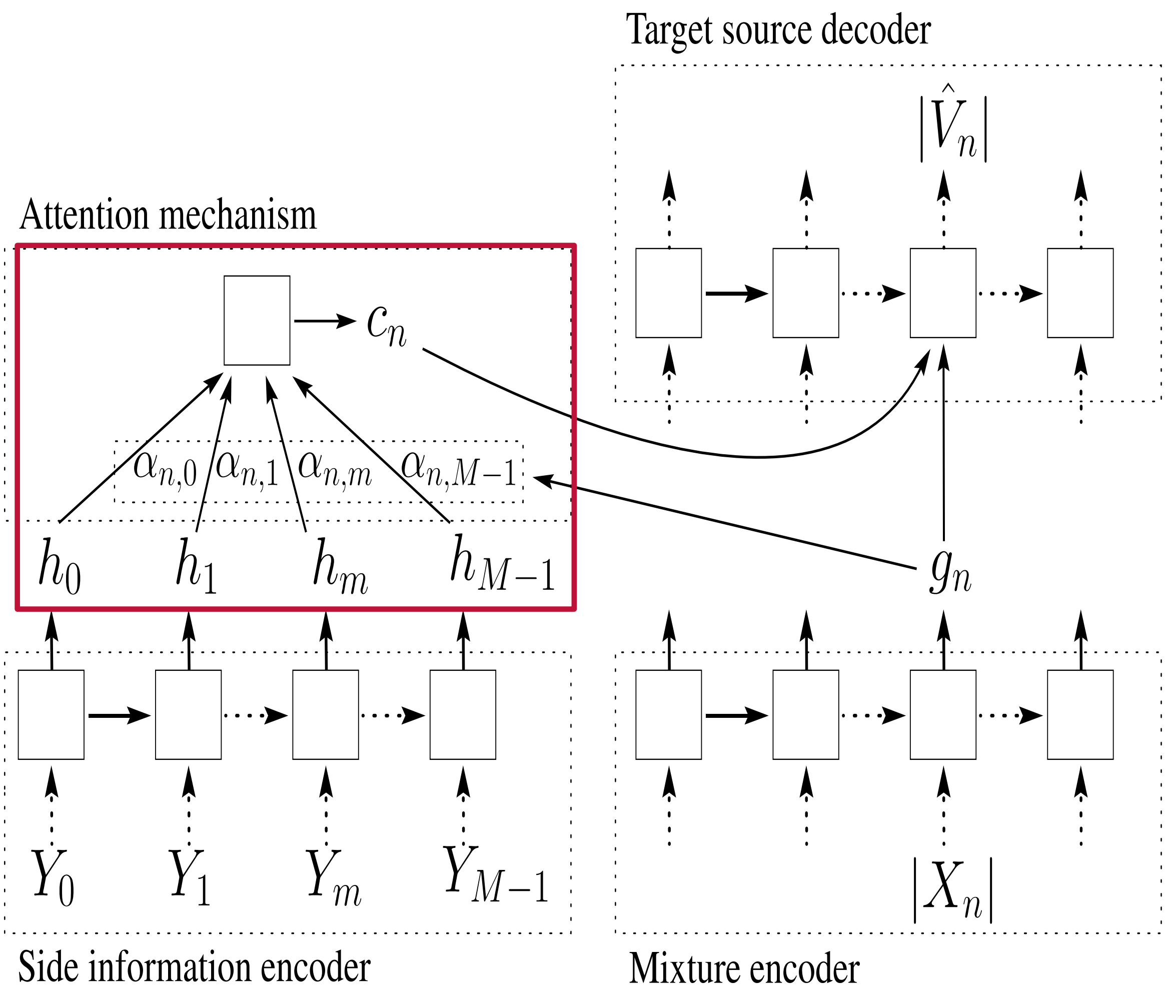


$$s_{n,m} = g_n^\top W h_m$$

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

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

Proposed Model: Learn to Align and Separate Jointly



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

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

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

Some architecture details

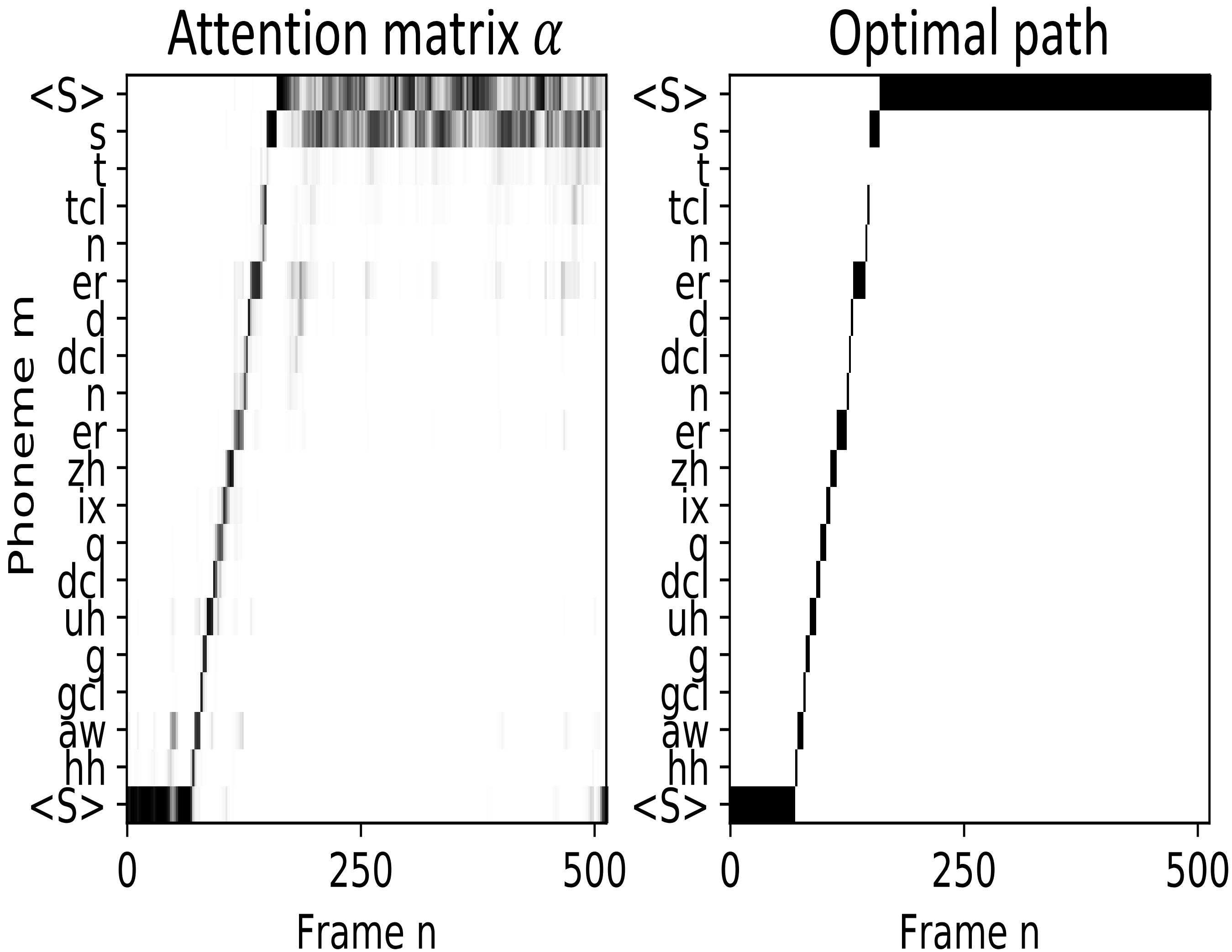
- **Input:**
  - Size side information $\neq$ size mixture audio input
- **Encoders:**
  - The mixture encoder is a two-layer bidirectionnal recurrent Neural Network with LSTM cells
  - Side information encoder is also a 2 layer BDRNN with LSTM
- **Decoder**
  - A first fully connected layer computes the hidden representation

$$q_n^{(1)} = \tanh(W_1[c_n, g_n] + b_1)$$

  - A two layers deep BRNN with LSTM cells (same as encoders) computes the hidden representation $q_n^{(2)}$
  - a fully connected layer with ReLU activation computes the estimation:

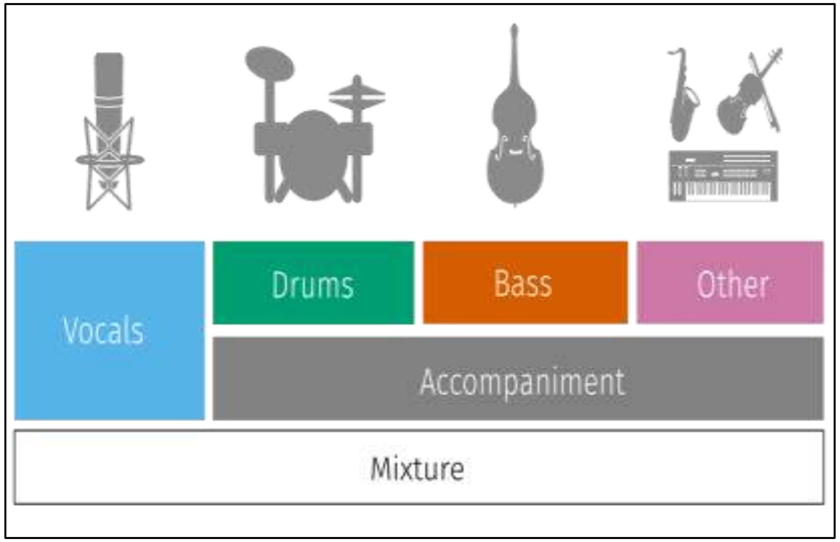
$$|\hat{V}_n| = \max(0, W_2 q_n^{(2)} + b_2)$$

Retrieving Phoneme Onsets from Attention Weights with Dynamic Time Warping (DTW)



Training Data: Mixtures, Clean Vocals, and Lyrics Transcripts

MUSDB18 corpus



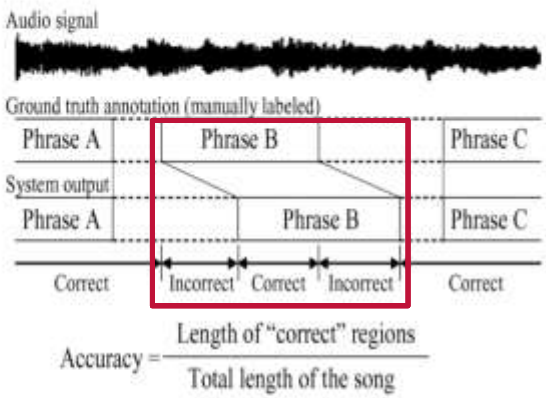
Coming soon: MUSDB lyrics extension

- 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

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
    - Solo singing recordings (1-3 minutes length)
    - **Accurate** phoneme transcripts with onsets
  - Mixed with MUSDB accompaniments
- **Baseline:** Montreal Forced Aligner
- **Metric:**



Method	PCAS [%]	SNR
ours	85.94	solo singing
baseline	77.94	solo 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

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.
McAuliffe, Michael, et al. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." *Interspeech*. 2017.
Fujihara, Hiromasa, et al. "LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics." *IEEE Journal of Selected Topics in Signal Processing,* 2011.

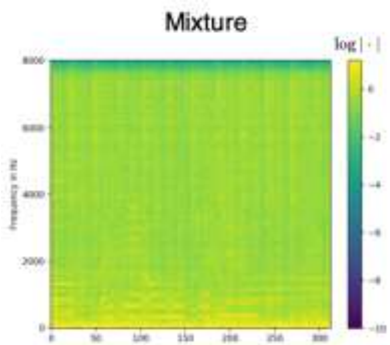
Results: Text-Informed Singing Voice Separation

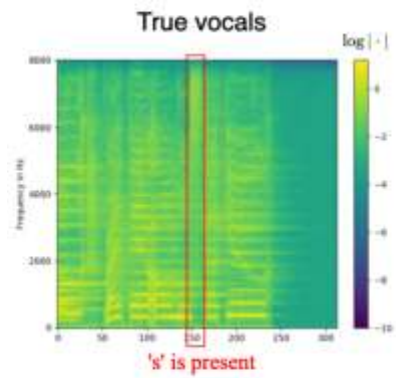
- **Test data:** MUSDB18 (only English songs)
- No improvement through text over baseline with **joint approach**
- **Improvements** through text in **sequential approach**:

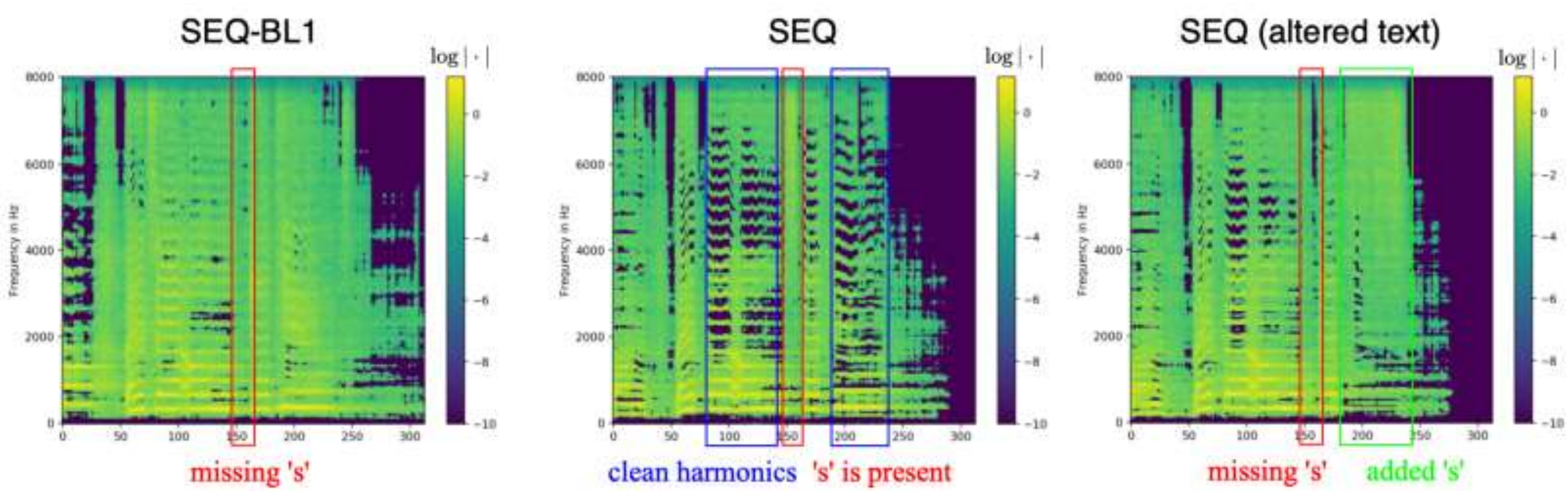
Side Info	1 singer			2+ sing. 1 phon.			2+ sing. 2+ phon.		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
constant	4.77	9.51	7.15	4.93	9.38	6.85	4.19	9.05	5.82
voice activity	4.74	9.17	6.83	4.55	9.14	6.94	3.75	8.62	5.70
phonemes	5.08	10.41	6.82	4.89	10.21	6.71	3.85	9.82	5.03

Evaluation scores in dB. Median over evaluation frames.

A short demo (1)







A short demo (2): White noise as input

- **The model SEQ, which was trained with aligned phonemes as side information, shapes white noise inputs according to the given phoneme information.**

White noise input	Text input	Output	True vocal
	> B AH T > SH IY Z > T UW > B L AY N D > T UW > S IY > IH N > M AY > K AA R > (but she's too blind to see in my car)		
	> T EY K > AE N > AE P AH L > P L IY Z > F R AH M > DH AH > F R AH N T > R OW > (take an apple please from the front row)		

Listening Examples

https://schufo.github.io/plla_tisvs/

Music Style Transformation using Sequence-to-Sequence Models

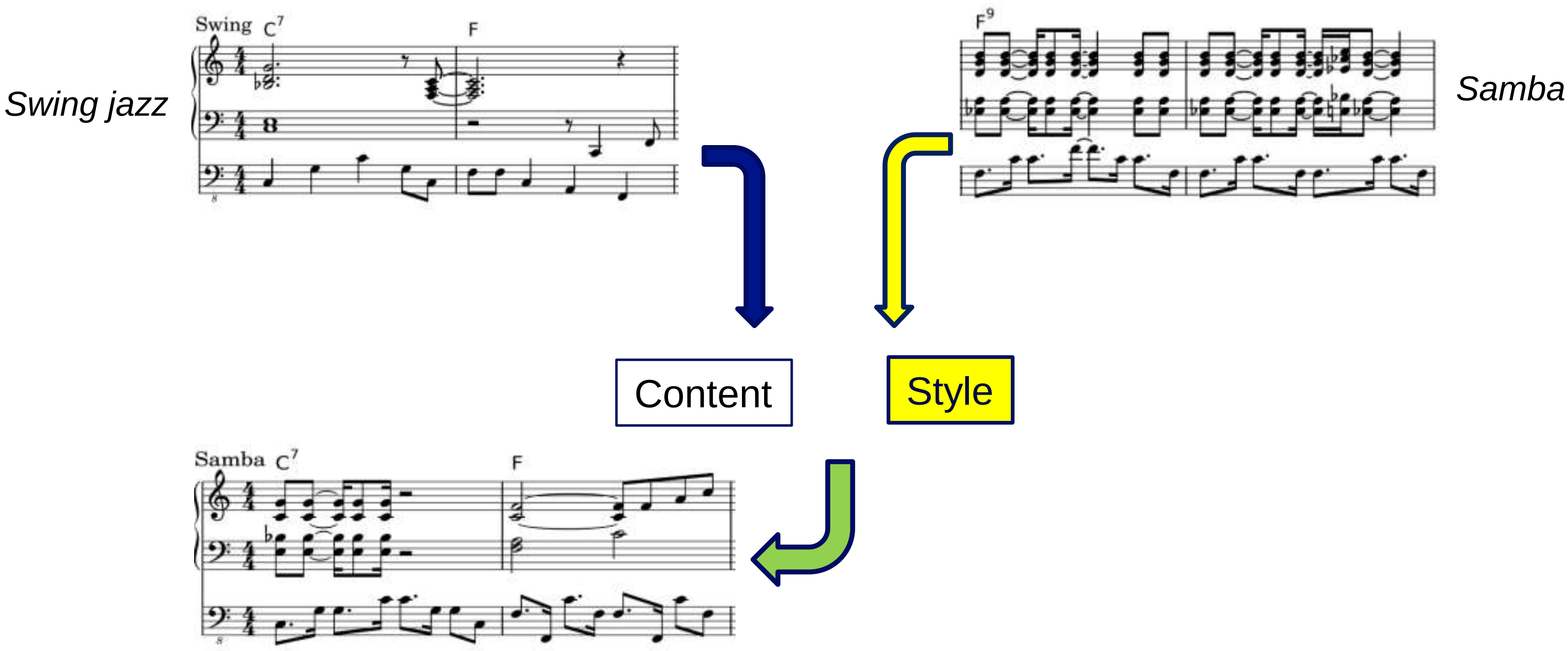
Ondrej Cifka, Umut Simsekli, Gaël Richard

MIPFrontiers

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.

Music style transfer

- … Or playing a given music file in the style of another music excerpt.



Ondrej Cifka, Umut Simsekli, Gaël Richard, “Groove2Groove: One-Shot Music Style Transfer with Supervision from Synthetic Data”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, (preprint) accepted for publication, 2020

Sound examples at : *https://groove2groove.telecom-paris.fr*

Analogy with Image

content: photo

+

style: *Starry Night* (van Gogh)

→

L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutionalneural networks. In CVPR, 2016.

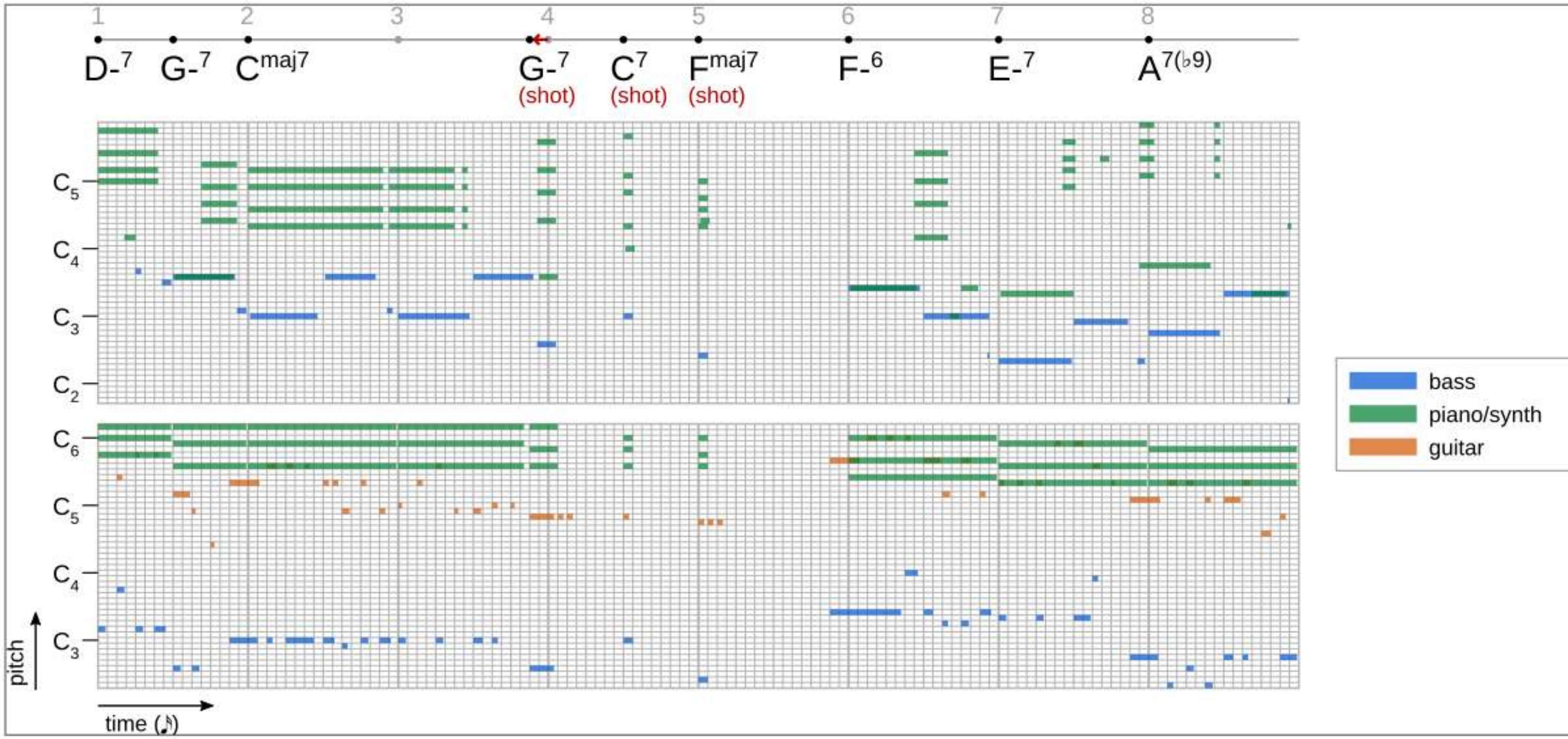
Music translation and style transfer

- **Music style translation**
  - To convert an input song to a target style known in advance
- **Music style transfer**
  - from a « content » song A and a « style » song B, to produce the song A in the style of B
- **Our interest:**
  - Convert an accompaniment (multiple tracks) to a different style, preserve harmonic structure (content);
  - Following a supervised approach; based on synthetic parallel data generated for this purpose

O. Cífka, U. Şimşekli, and G. Richard. Supervised symbolic music style translation using synthetic data. In *ISMIR*, 2019.
O. Cífka, U. Şimşekli, and G. Richard. Groove2Groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, 2020.

Music translation

- **Starting from chord charts, use Band-in-a-Box (BIAB) software to generate accompaniments in 70 different styles**



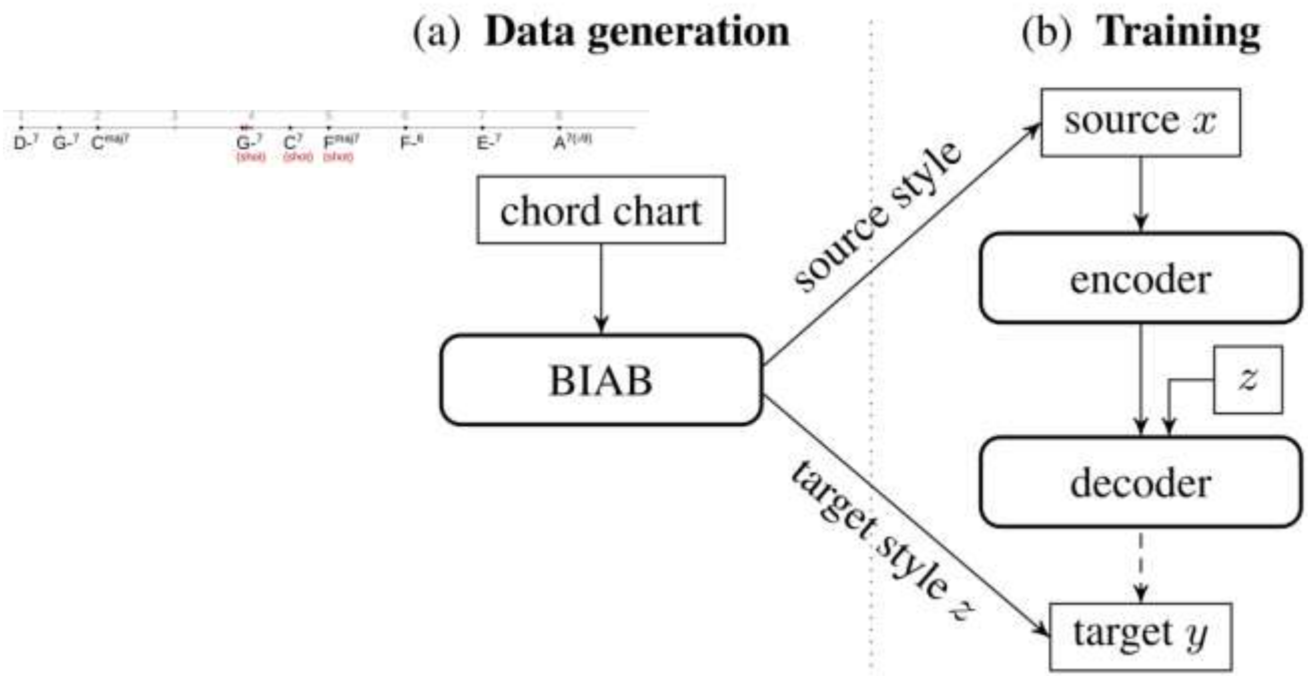
Droits d'usage autorisé

TELECOM Paris

IP PARIS

Music translation

- **Data and Training principle**
  - The model is trained to predict the target-style segment y given a source segment x and the target style z

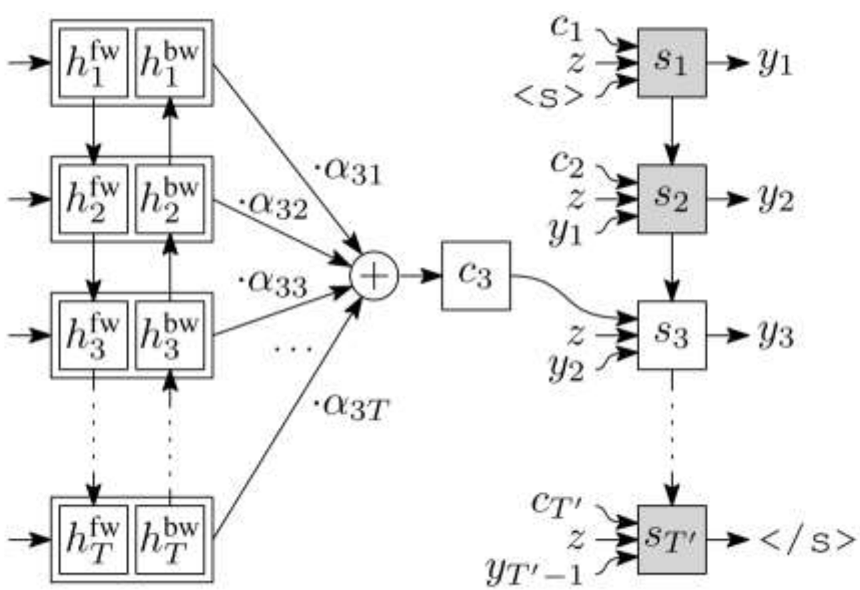


Music translation

- **A few details on the decoder**

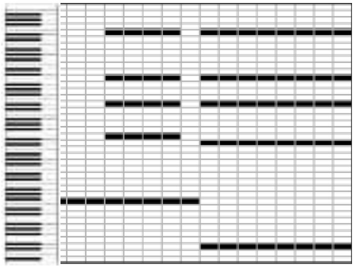
$$s_i = \mathrm{GRU}([c_i, W^{\mathrm{s}} z, W^{\mathrm{e}} y_{i-1}], s_{i-1}),$$

  - c_i is the context vector
  - $W^s z$ is the style (weighted by corresponding embedding)
  - $W^e y_{i-1}$ is the previous output event (weighted by corresponding embedding)
  - s_{i-1} is the previous state



Music translation: model

- **Based on seq2seq from machine translation encoder (K. Cho &al.)**
  - 2 layer CNN
  - followed by a bidirectional RNN with a gated recurrent unit (GRU)
- **Decoder:**
  - RNN with attention
- **Input:**
  - piano roll matrix



- **output:**
  - token sequence encoding MIDI events trained on pairs (*x*,*y*); *z* is the style of *y* one model per instrument (bass, piano)

```
NoteOn(50) TimeShift(9) NoteOn(60) NoteOn(65)
NoteOn(69) NoteOn(76) TimeShift(12) NoteOff(60)
NoteOff(65) NoteOff(69) NoteOff(76) TimeShift(3)
```

Kyunghyun Cho &al.. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP, 2014.

Evaluation /results

Evaluation metrics

- *Content preservation* :
  - measures harmonic similarity between the input and the output
  - column-wise cosine similarity of smoothed "chromagram"
- *Style fit metric (proposed)*
  - collect statistics of musical events (note pitch, onset time, duration, velocity) → style profiles
    - 2D histograms: time-pitch, onset-duration, …
  - compute cosine similarity between output and reference

W.T. Lu and L. Su. Transferring the style of homophonic music using recurrent neural networks and autoregressive models. In *ISMIR*, 2018.

Evaluation /results

- **"Almost perfect" results, especially on style fit metrics**
  - the network is **able to imitate the training styles**
  - correctly **follows the harmony** of the input

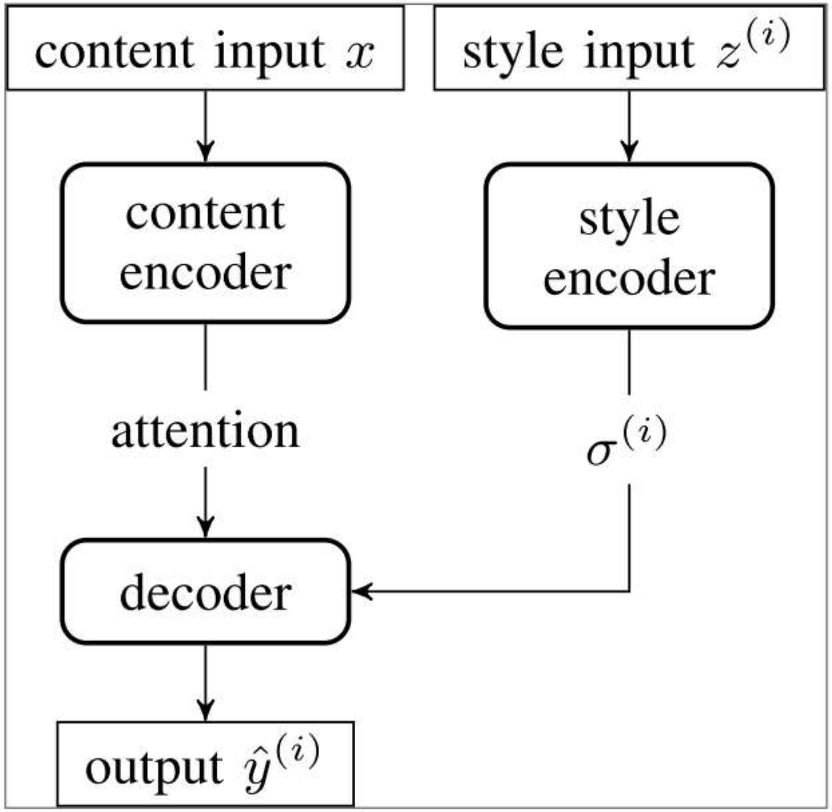
- **Generalizes to arbitrary MIDI inputs**

- **Main limitation: cannot generalize to new target styles**

→ **Interest for One-shot style transfer**

One-shot style transfer

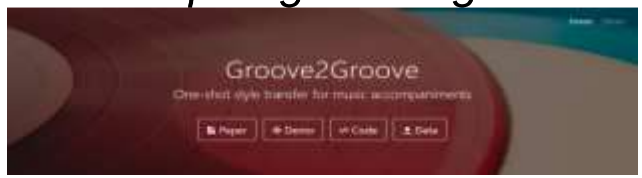
- Extends the style translation model by adding the style encoder
- Common model for all instruments trained on triplets $(x, y^{(i)}, z^{(i)})$
- Data contains 3k different styles; generated so that train, val & test sections use disjoint sets of styles

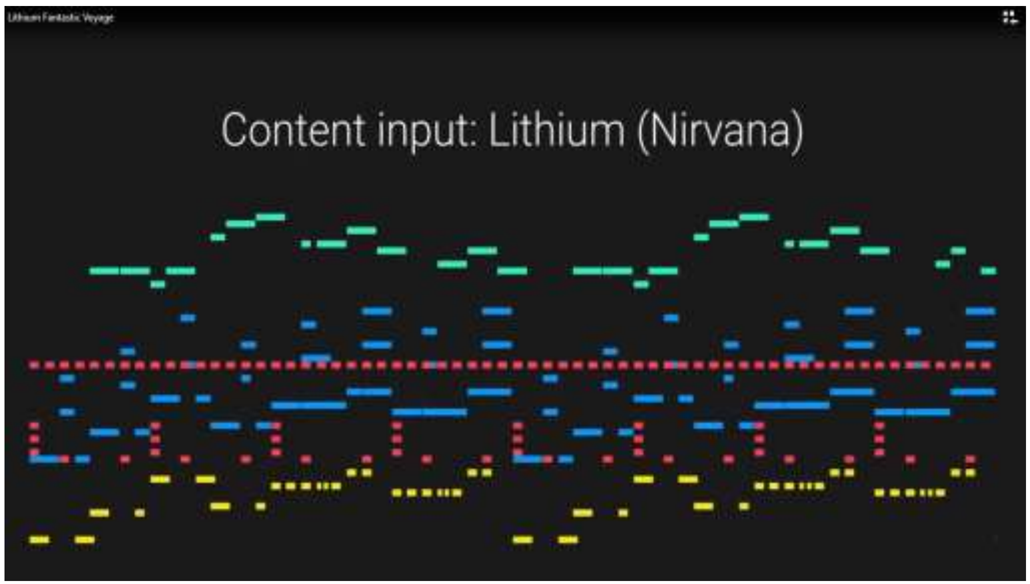


One-shot music style transfer

- **A short demo**

(more sound examples at : https://groove2groove.telecom-paris.fr)





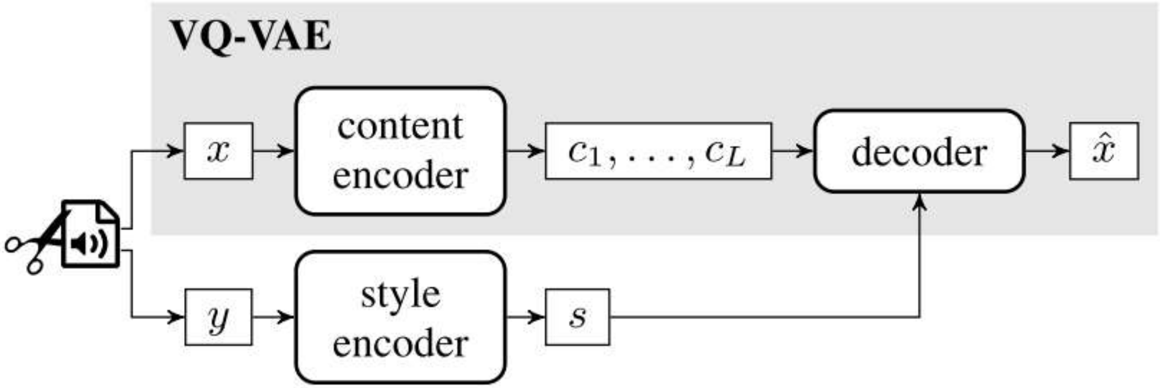
Ondrej Cifka, Umut Simsekli, Gaël Richard, “Groove2Groove: One-Shot Music Style Transfer with Supervision from Synthetic Data”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, 2020

Sound examples at : *https://groove2groove.telecom-paris.fr*

Variation: one-shot timbre transfer

Demo at: https://cifkao.github.io/ss-vq-vae/

Use of VQ-VAE



- **A discrete representation** is used for content and the model is trained to reconstruct the content input, x;
- The **output of our style encoder** is a single continuous-valued embedding vector s.
- Use of a **simple self-supervised learning strategy** (i.e. x and y are different segments of the same audio recording) => (goal: style encoder only encodes style and is content-independent),

O. Cifka A. Ozerov, U. Simsekli, G. Richard, Self-Supervised VQ-VAE For One-Shot Music Style Transfer, in Proc. ICASSP 2021.

Conclusion

- **Deep neural Network for speech, audio and music is very active**

- **For audio**
  - A clear interest for architectures which are capable of modelling time series (e.g. context)
  - A clear interest for RNN, GANs, Attention mechanisms, Transformers,….
  - A trend towards more frugality, hybrid models mixing « signal knowledge » and power of Deep learning.