Estimating music and speech descriptions using deep learning

Alice Cohen-Hadria
Analysis Synthesis Team
Supervision: Geoffroy Peeters

October 28th, 2019

Jury : Simon Dixon - Reviewer
Emmanuel Vincent – Reviewer
Carlos Agon - Examiner
Juan Pablo Bello – Examiner
Isabelle Bloch – Examiner
Axel Röbel – Examiner
Jimena Royo-Letelier - Examiner

1.

Introduction

Audio Content Analysis

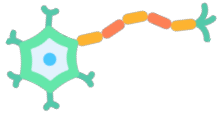
Goal: describing/processing/separating the content of:

- Speech,
- Music (MIR),
- Environmental sound (DCASE).

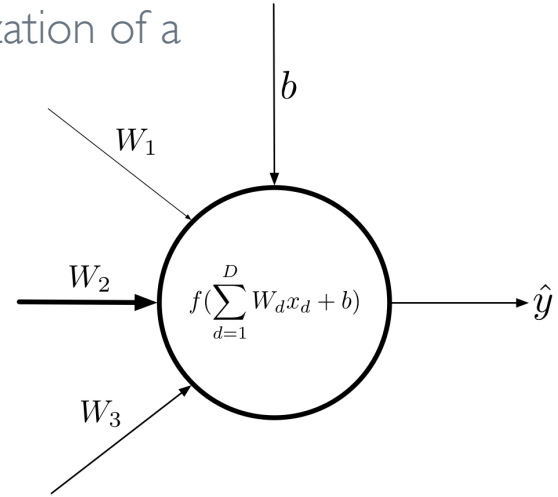
How to solve these problems?

- Traditional approaches:
  - Hand-crafted features + ML.
  - ICA, NMF, PLCA.
- Large breakthrough, thanks to **deep-learning.**

What is deep learning ?



- Based on neural networks.
- One neurone is mathematical simplification of a modelization of a biological neuron.



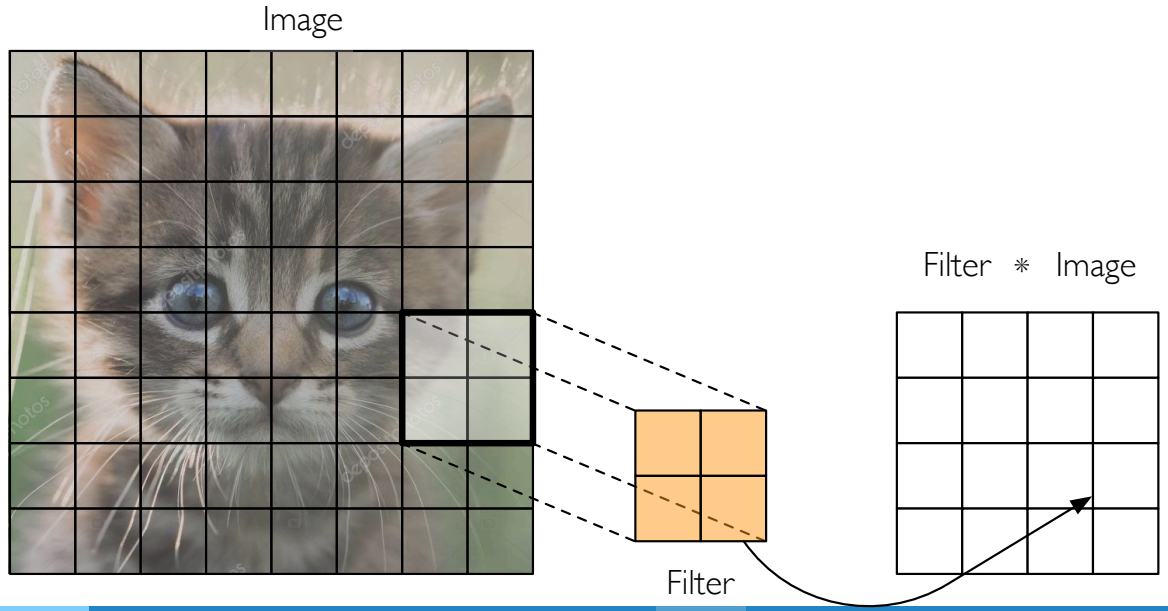
- In this work, focus on

  Convolutional Neural Network (ConvNet) [Fukushima,1980] [LeCun et al., 1989].

Convolutional Neural Networks (ConvNet) - I

Designed to analyze images

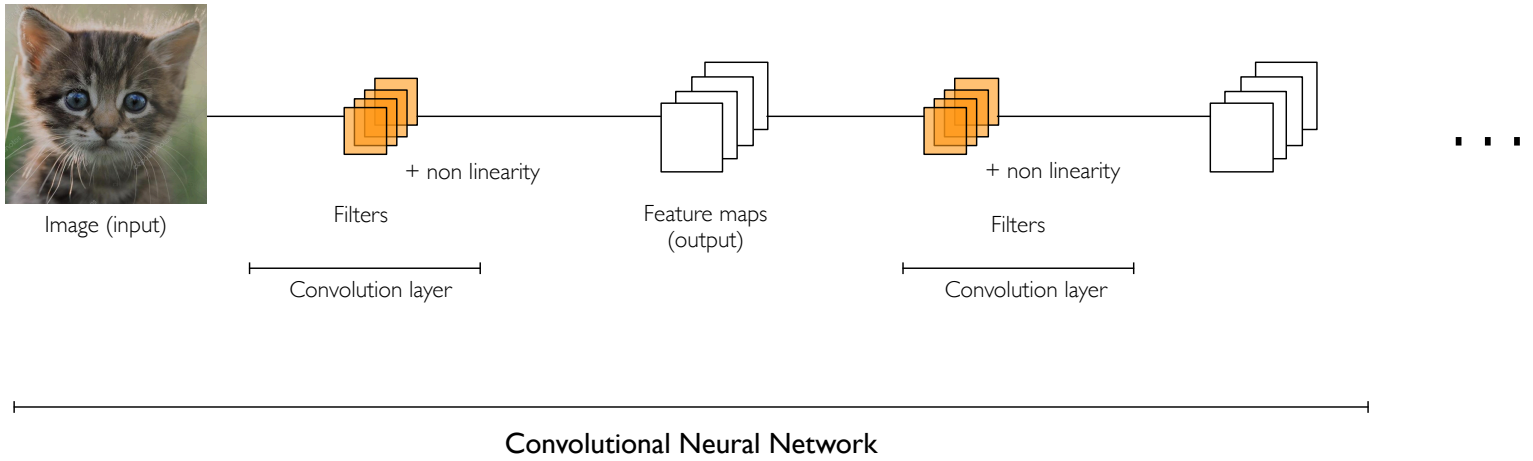
The base operation of ConvNet is **convolution** of filters along the images:



One filter represents a type of visual feature (straight lines …). It activates in all the portion of the studied image similar to this feature.

Convolutional Neural Networks (ConvNet) - II

To form a network: connect the outputs of one layer to the inputs of the next.



Convolutional Neural Networks (ConvNet) – III

- We use here supervised learning.
- Supervised learning ? pairs of (input, output) presented to the network.
  - Need for labeled data.
- Training ?
  - Iteratively modifying filters' elements to reduce prediction error (loss function).
  - Gradient descent like (backpropagation).

- ConvNet ?
  - Led to huge improvements in computer vision: image classification, image segmentation …

Problems

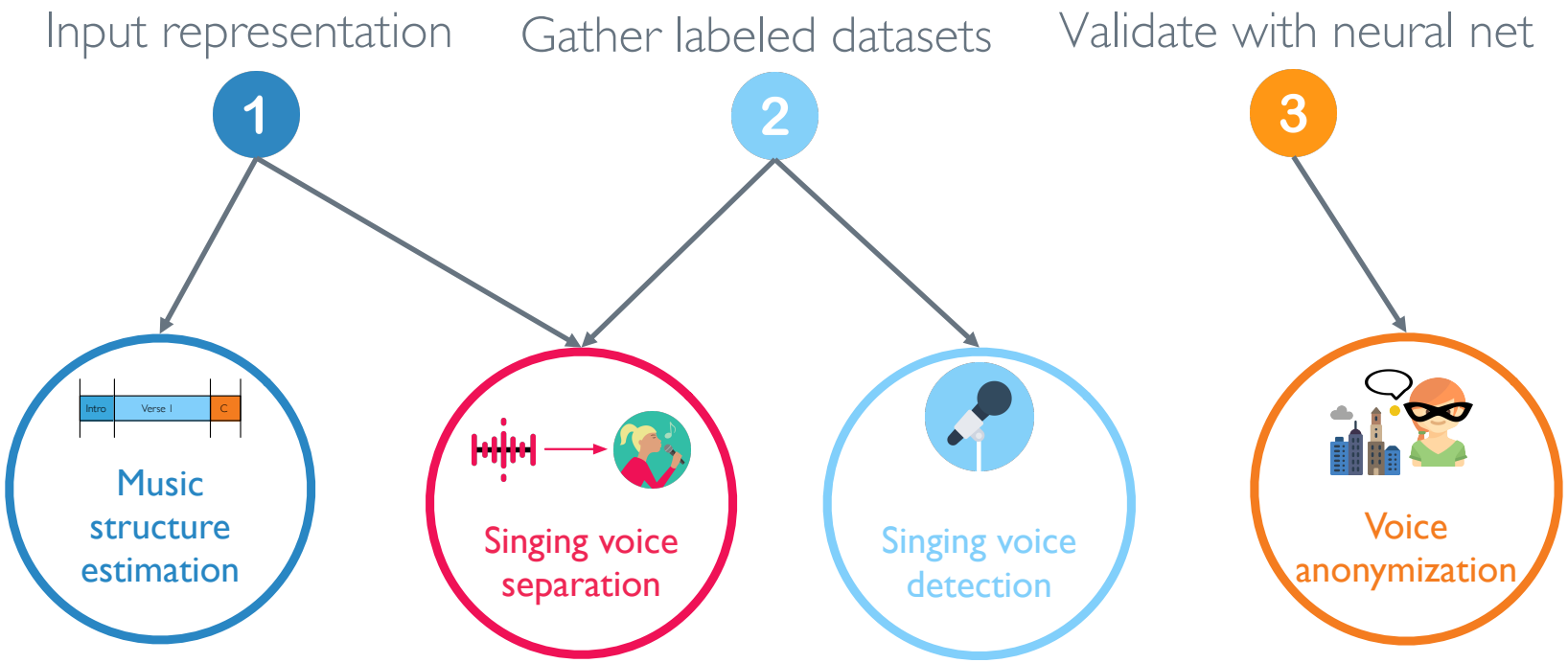
For sound and music:

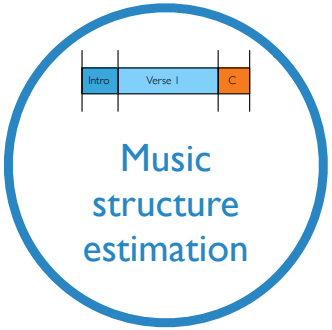
1. Image is two D. Sound is one D.
   - Waveform ([Sainath, 2015]) or prior knowledge?

2. Computer vision field has large datasets.
   - How to gather large labeled datasets for audio and music ?

3. ConvNet are powerful tools.
   - Can we use ConvNet not only for solving a problem but also to validate the solution found ?
   - Useful for validation.

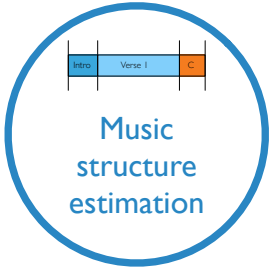
To study those problems: 4 tasks





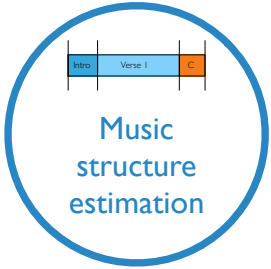
Tasks presentation

What is music boundary estimation?



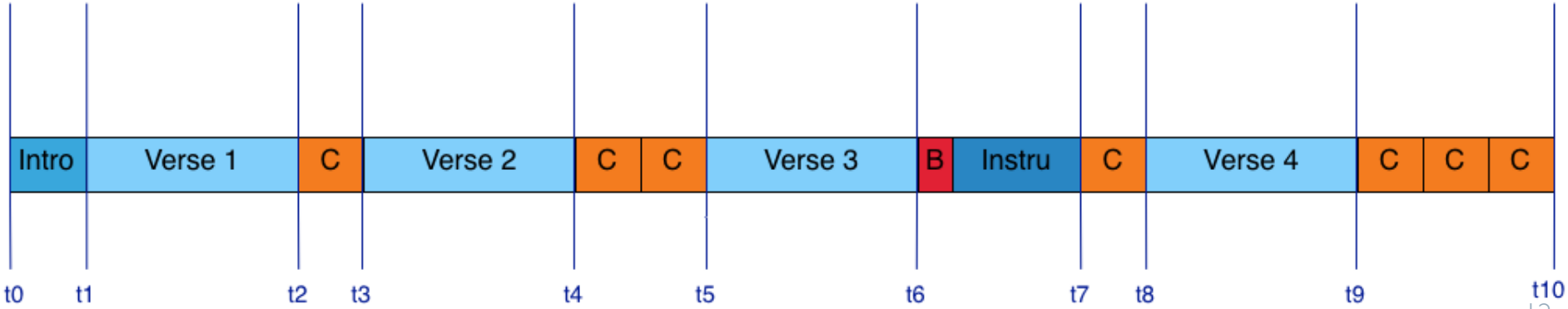
Estimating automatically the temporal structure of a music track by analyzing the characteristics of its audio signal over time.

What is music boundary estimation?

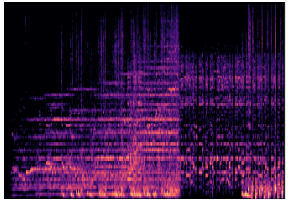


Estimating automatically the temporal structure of a music track by analyzing the characteristics of its audio signal over time.

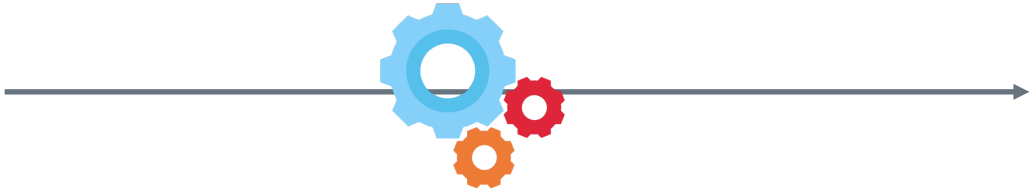
Estimating the boundaries between those segments.

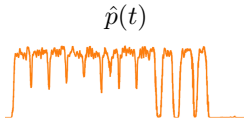


Singing voice detection



Audio track





Singing voice
probability

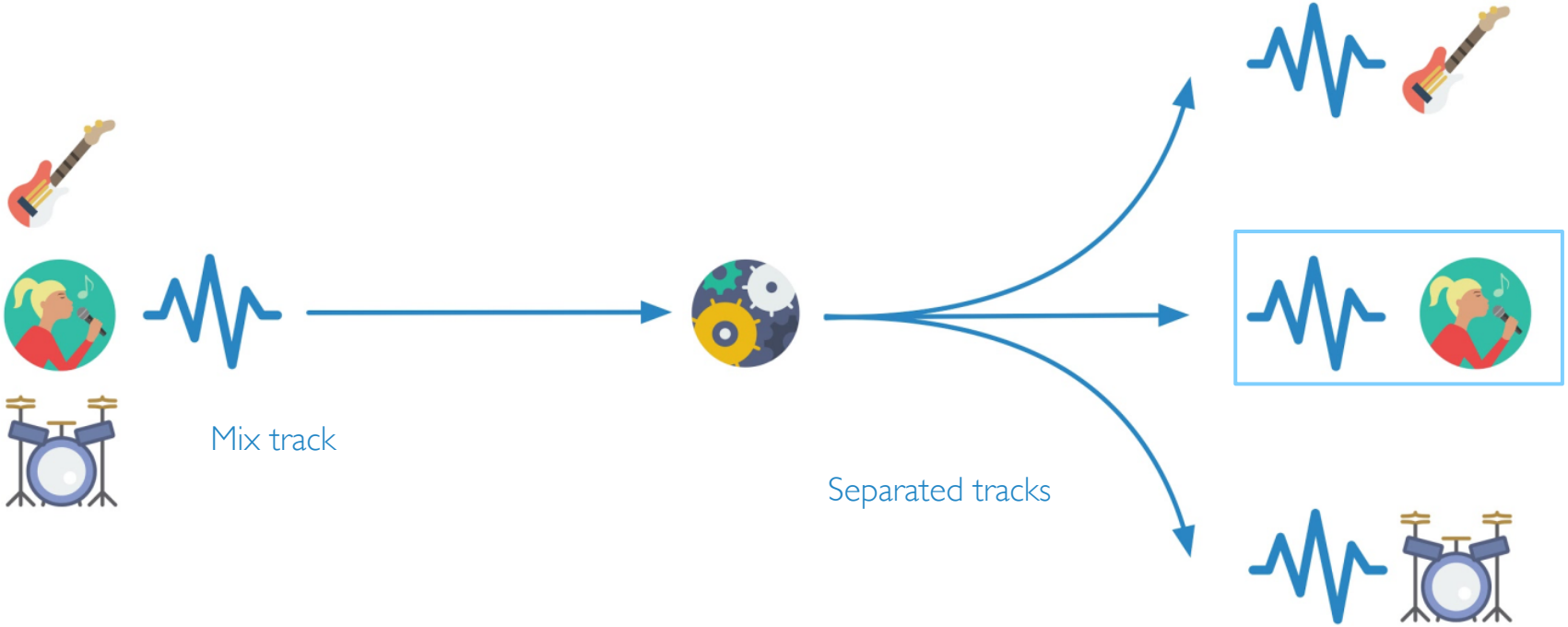
From an audio track to singing voice probability.

Usefull as preprocessing task, or to do analysis.

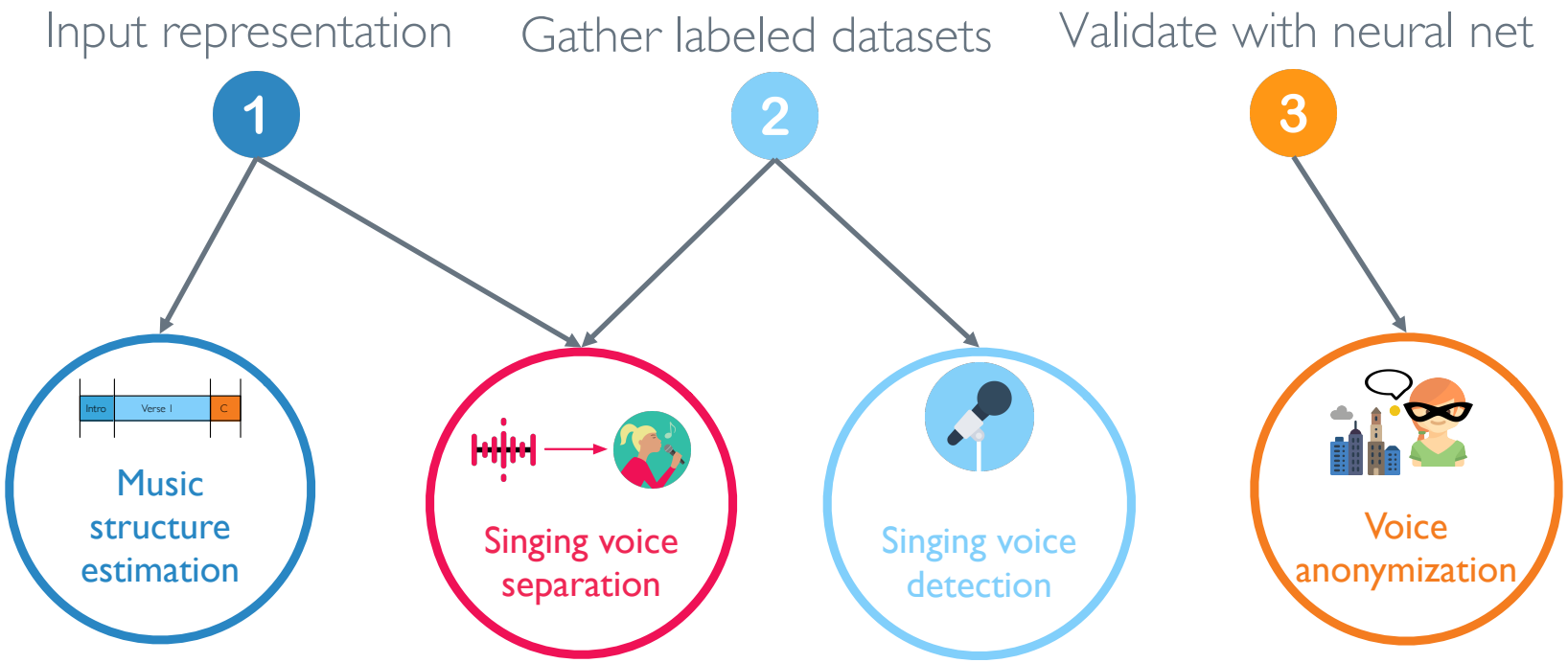
What is singing voice separation ?

Mix track

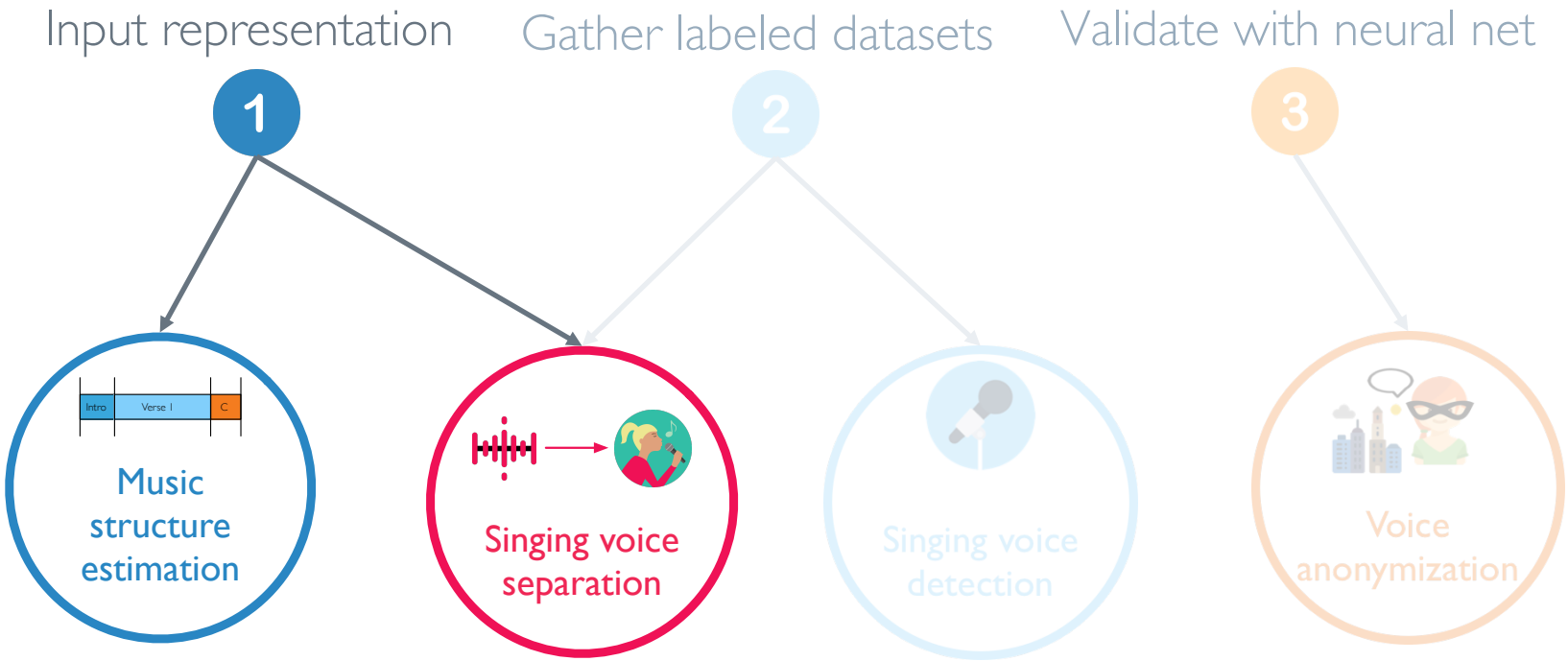
What is singing voice separation ?



To study those problems: 4 tasks



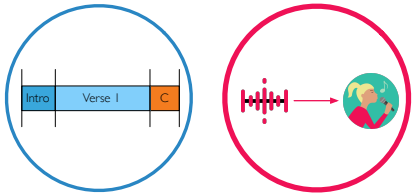
To study those problems: 4 tasks



II.

Choosing the correct input respresentation

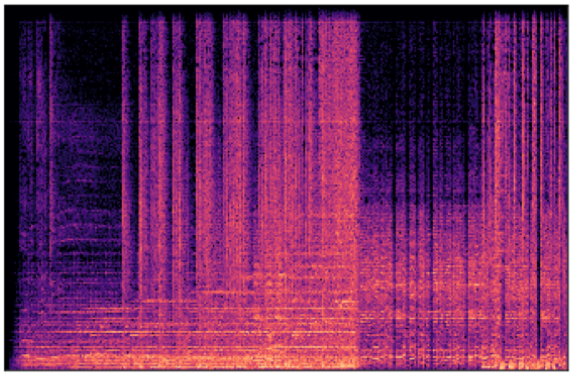
Use cases: music boundaries estimation and singing voice separation
Work published in [Cohen-Hadria and Peeters, 2017] and [Cohen-Hadria et al 2019a]



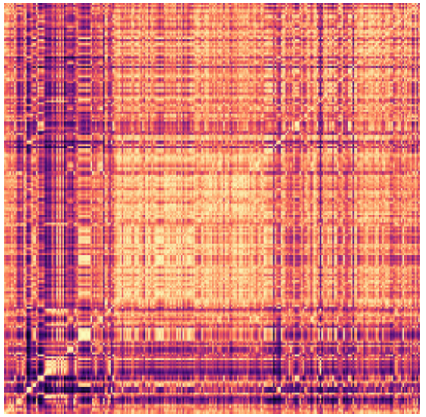
Input representation for ConvNet

- Sound is 1D.

- But 2D representation exists. For sound:



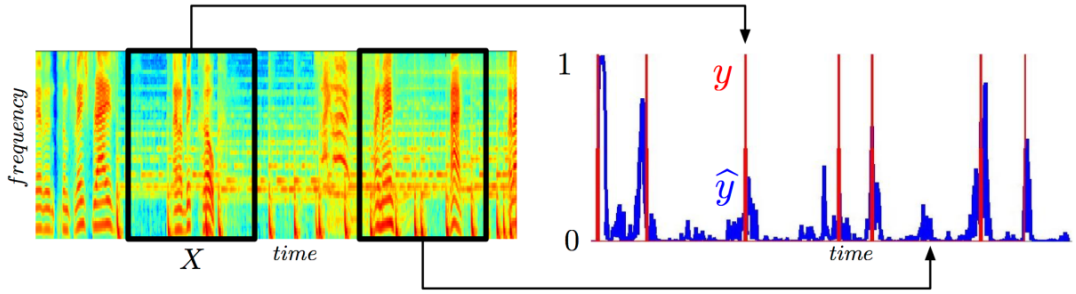
Spectrogram or Mel Spectrogram



Self Similarity Matrix (SSM)

Music boundaries estimation

ConvNet for music boundaries estimation



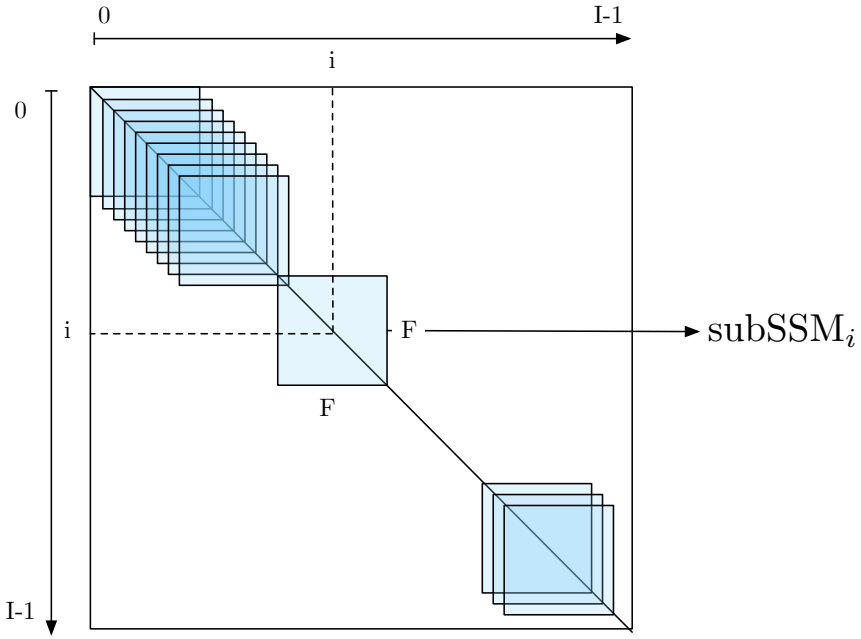
Input representation previously used:

- Mel Spectrogram (MLS) [Ullrich et al., 2014] .
- or Combination of MLS and Lag Matrix [Grill and Schluëter, 2015].
  - But Lag Matrix is not invariant over time.

Which kind of representation should we choose for music boundaries estimation ?

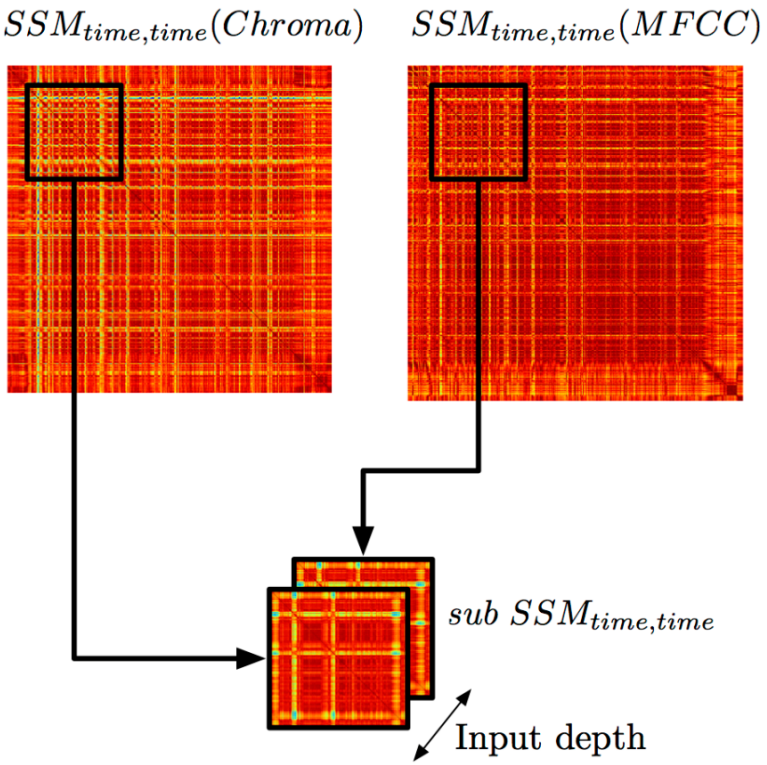
Our proposed input representation: SSM

- Use square-sub-matrices centered on the main diagonal of a Self-Similarity-Matrix time-time as input.
- Already used by [Foote, 2000] or by [Kaiser and Peeters, 2013]. Provides sharper edges at the beginning and ending of segments.



Our proposed input representation: SSM stacked

- RGB representation for colored image
- Stacked SSM as input of a convolutional network.
- Two points of view. MFCC and Chroma

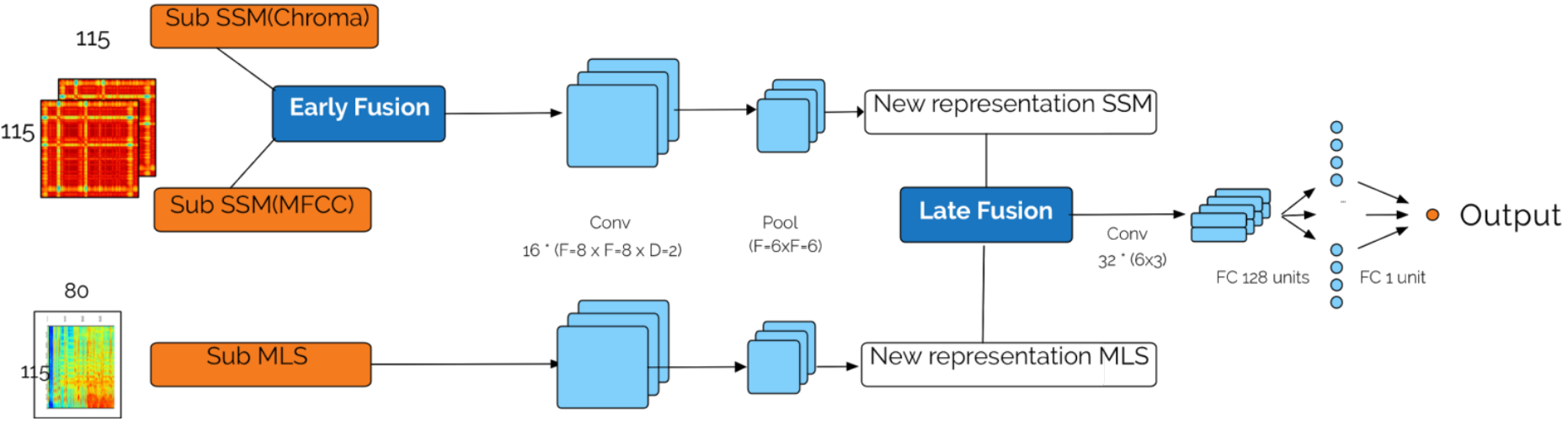


Using input depth - Architecture used

Late fusion of MLS and SSM. Two sub-network with inputs:

- SubNetwork 1: Uses SSM_Single (MFCC or Chroma) or SSM_stacked.
- SubNetwork 2: Uses MLS.
- ConvNet3 for the fusion of the two representations.

Using input depth - Architecture used



Experiments

Train models with the following input combinations:

- ① MLS + SSM_mfcc
- ② MLS + SSM_chroma
- ③ MLS + SSM_stacked

- MLS + SSM_lag MFCC [Grill and Schluëter, 2015]
  - ④ reimplemented • ⑤ published

Side result - reproductibility

Didn't reach state-of-the-art results

Possible reasons:

- We didn't have access to their code
- We didn't have access to their full training-set
- Important to share those for reproducibility

Model	F-M 0.5s (std)	AUC
④ [Grill and Schluëter, 2015] reimplemented	**0.246** (0.112)	0.774
⑤ [Grill and Schluëter, 2015] published	**0.523**	

Results – Lag ④ versus Time ① ②

Using the self-similarity matrix expressed in time ① ② rather than in lag ④ provides an improvement at ±0.5 s and ±3 s.

Model	F-M 0.5s (std)	AUC
① MLS + SSM_mfcc	**0.273** (0.132)	0.810
② MLS + SSM_chroma	0.270 (0.153)	0.800
④ [Grill and Schluëter, 2015] reimplemented	**0.246** (0.112)	0.774

Results – Single ① ② versus Stacked ③

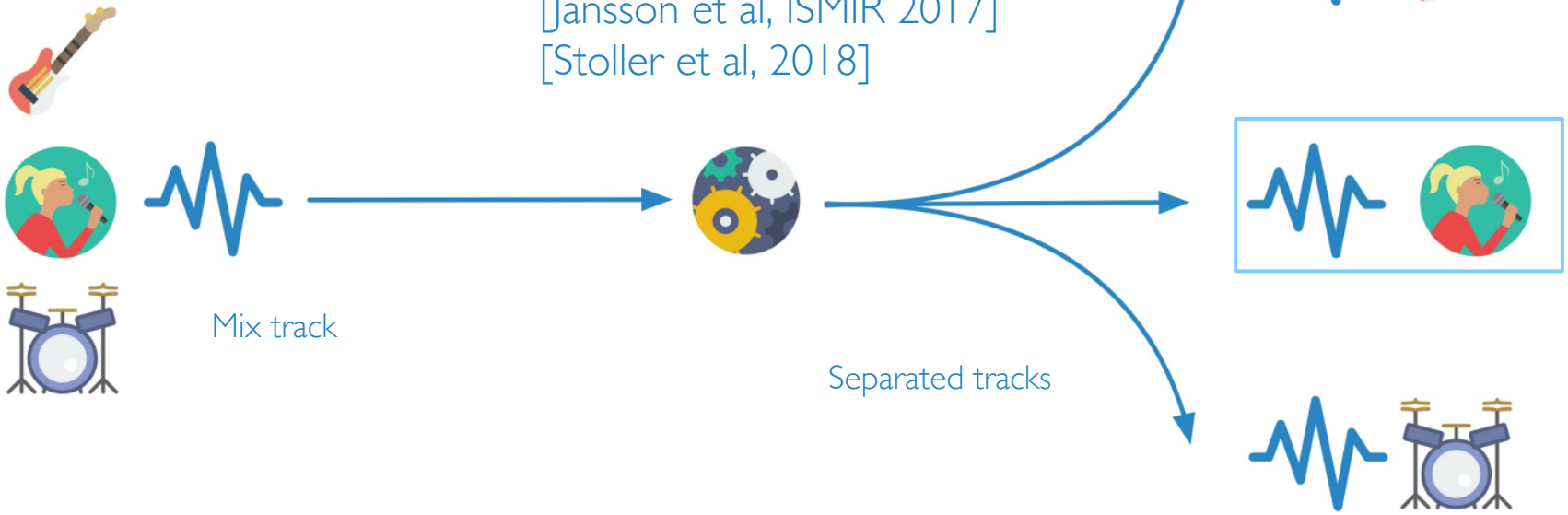
Using the depth of the input layer to combine the two SSM ③ allows us to increase the F-measure at ± 0.5 s. and ± 3 s.

Model	F-M 3s (std)	AUC
① MLS + SSM_mfcc	**0.551** (0.158)	0.946
② MLS + SSM_chroma	0,540 (0.153)	0.922
③ MLS + SSM_stacked	**0.629** (0.164)	0.930

Singing voice separation

State of the art

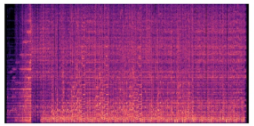
ConvNet :
[Jansson et al, ISMIR 2017]
[Stoller et al, 2018]



Mix track

Separated tracks

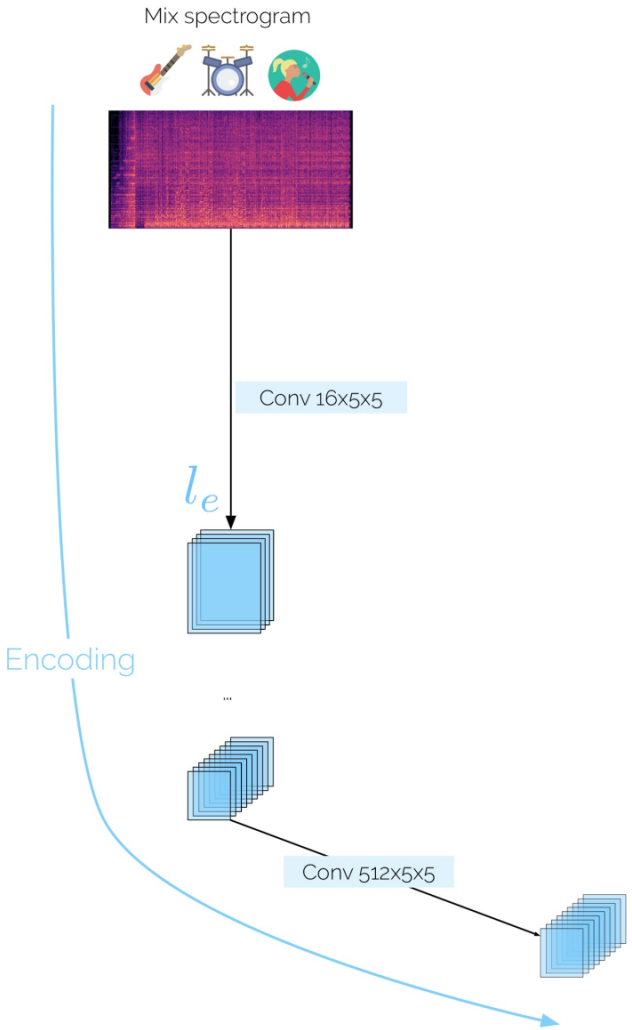
Mix spectrogram



U-Net

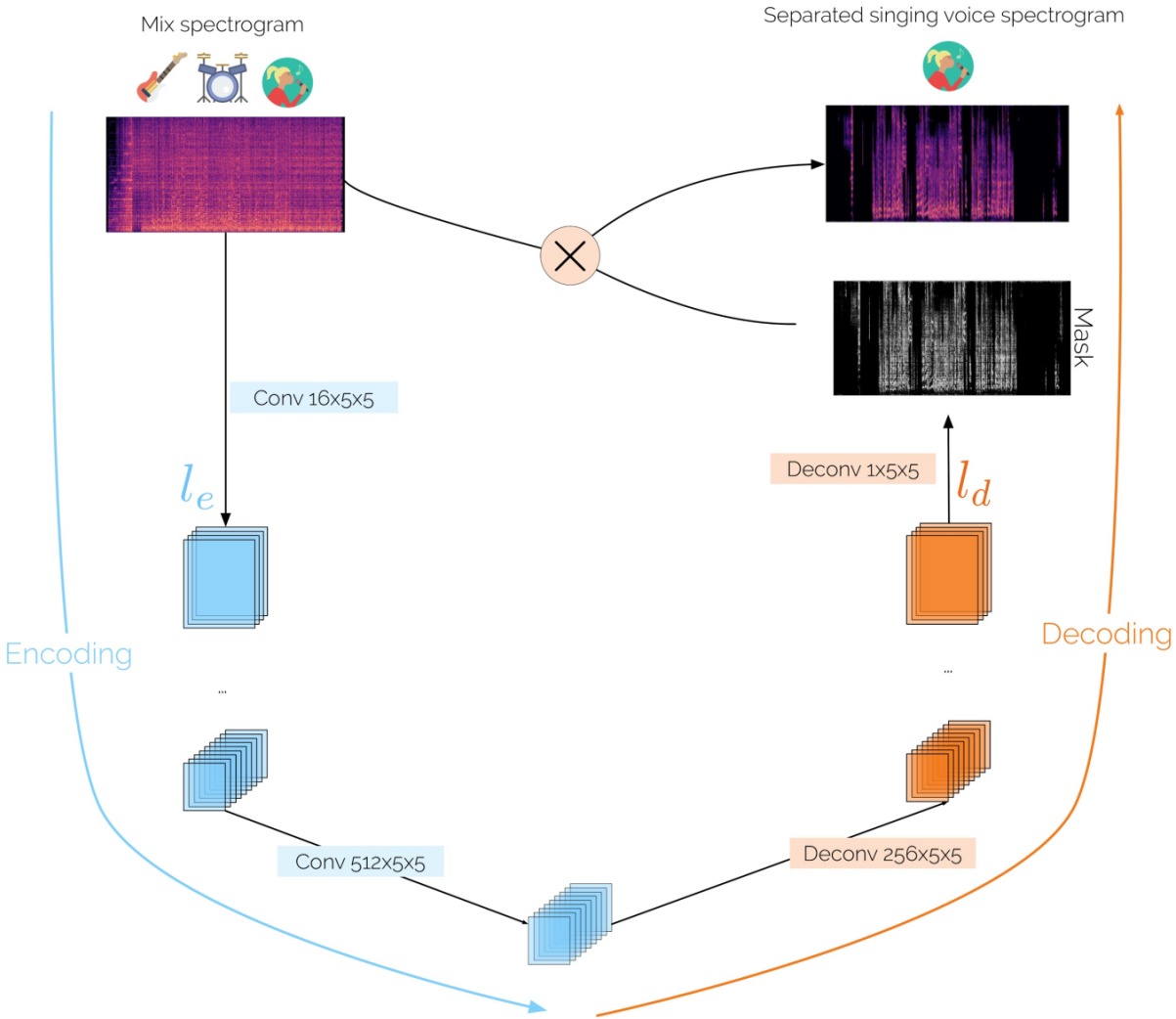
U-Net

Encoding-decoding scheme



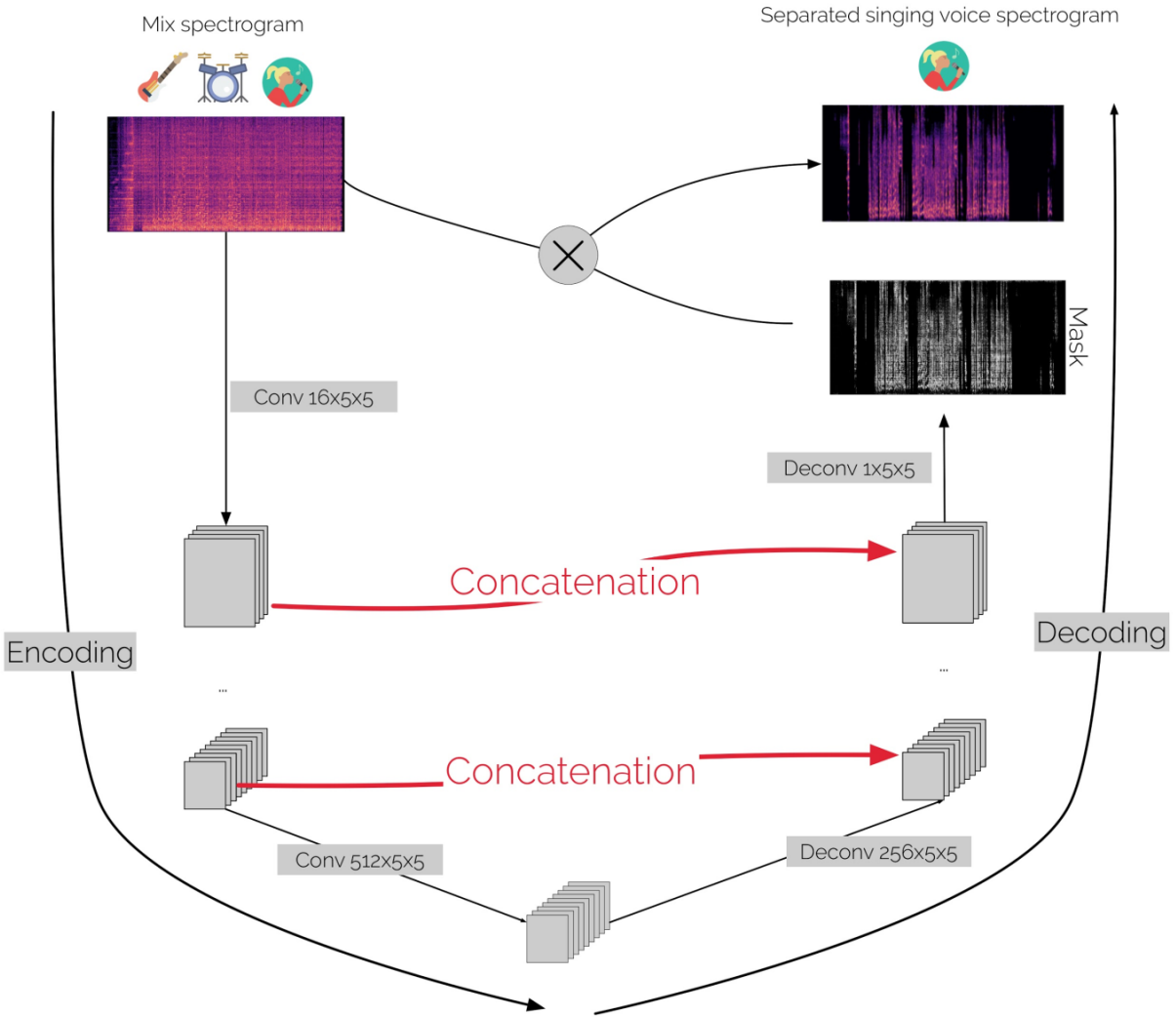
U-Net

Encoding-decoding scheme



U-Net

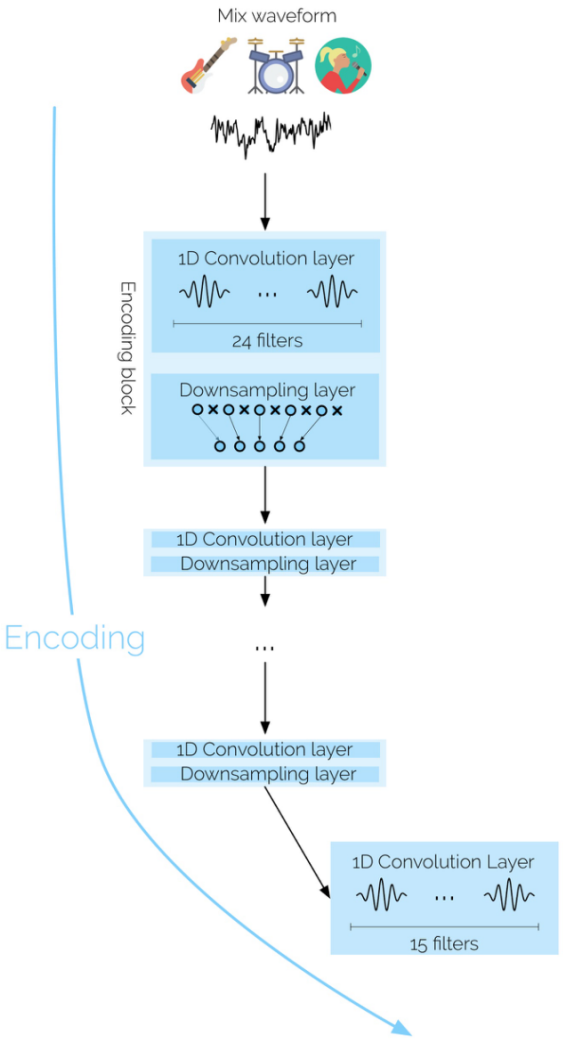
Skip Connections



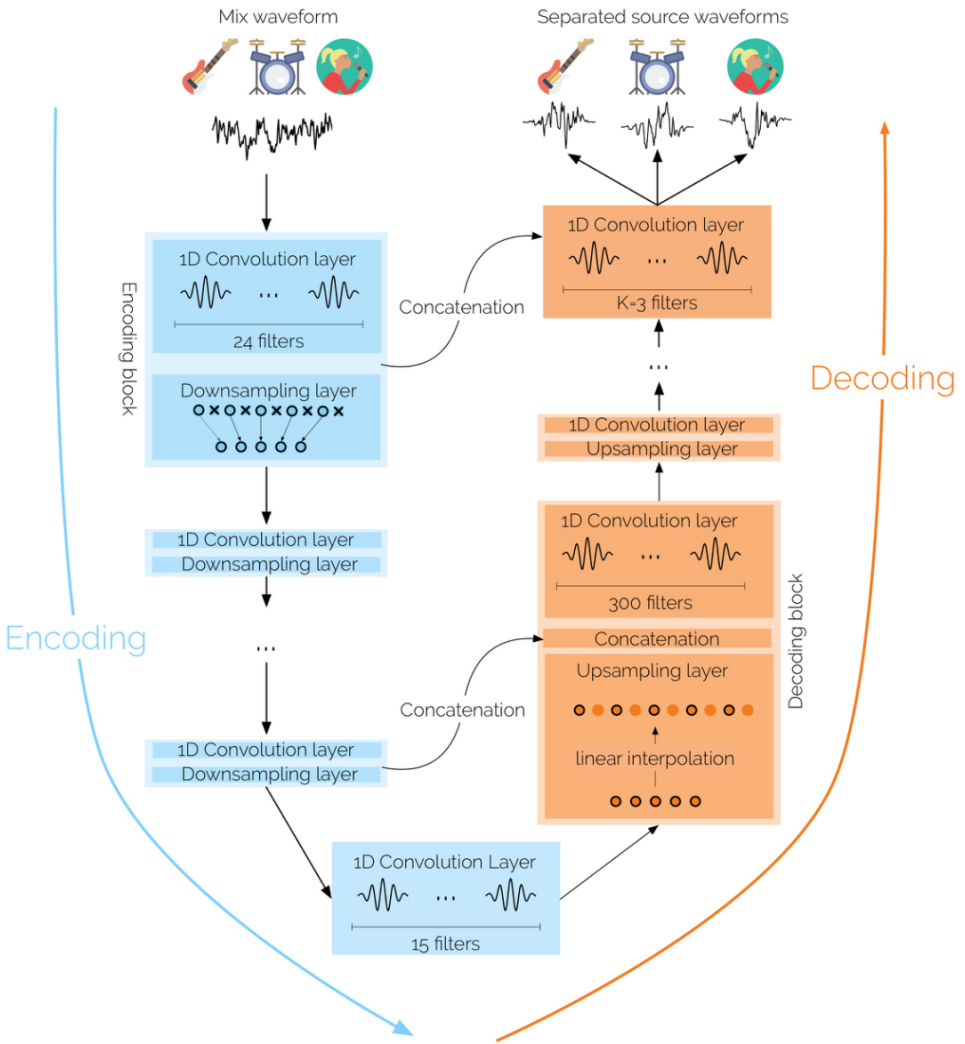
Wave-U-Net

- U-Net's adaptation with waveform as input.
- Presented on [Stoller et al, 2018].
- Adapted to be compared to U-Net.
  - new sampling rate and need of a large dataset to compare both models.

Wave-U-Net


Mix waveform
Encoding
Encoding block
1D Convolution layer
24 filters
Downsampling layer
1D Convolution layer
Downsampling layer
...
1D Convolution layer
Downsampling layer
1D Convolution Layer
15 filters

Wave-U-Net


Mix waveform
Separated source waveforms
1D Convolution layer
24 filters
Downsampling layer
Encoding block
1D Convolution layer
Downsampling layer
1D Convolution Layer
15 filters
Concatenation
1D Convolution layer
K=3 filters
1D Convolution layer
Upsampling layer
1D Convolution layer
300 filters
Concatenation
Upsampling layer
linear interpolation
Decoding block
Encoding
Decoding

Comparison U-Net Wave-U-Net

- Having more data: waveform better than spectrogram
- Not by much: interesting given the different representations.

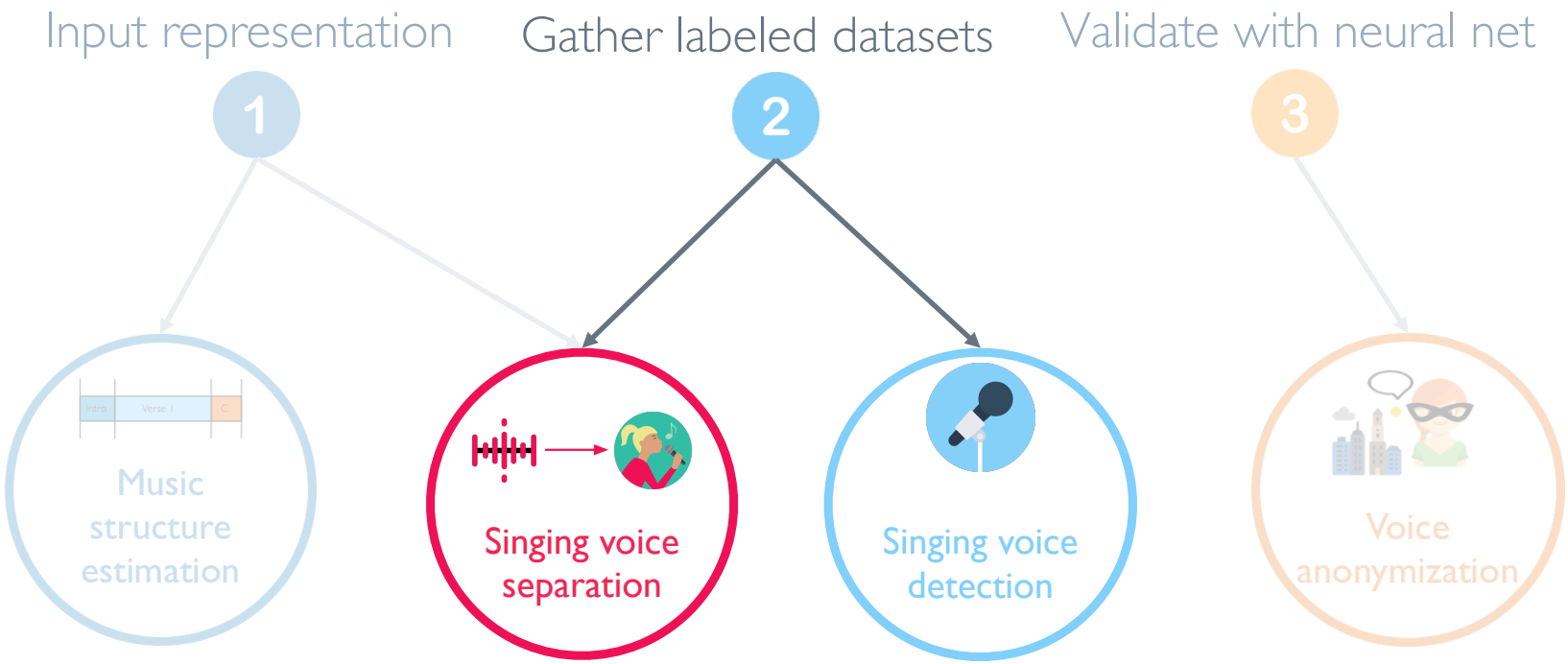
	SAR		SDR	
Data	U-Net	Wave-U-Net	U-Net	Wave-U-Net
Musdb **small**	**5.76**	5.52	**4.52**	4.09
Musdb **large**	6.40	**6.62**	5.30	**5.42**

Partial conclusion

- Using prior knowledge/signal processing representation is useful, when having few data:
  - Structure estimation.
  - Singing voice separation.

- When having large datasets, waveform can be enough of a representation.

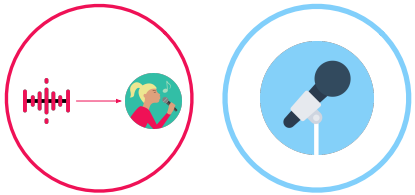
To study those problems: 4 tasks



III.

How to gather large amount of labeled data?

Use case: singing voice separation and detection
Work published in [Cohen-Hadria et. al, 2019a] and [Meseguer-Brocal, Cohen-Hadria & Peeters, 2018]



How to gather large amount of labeled data?

- Two strategies presented here :
  - Hard to label data : **data augmentation.**
  - Ressources online that can be used: **Teacher/Student paradigm.**

Singing voice separation & Data Augmentation

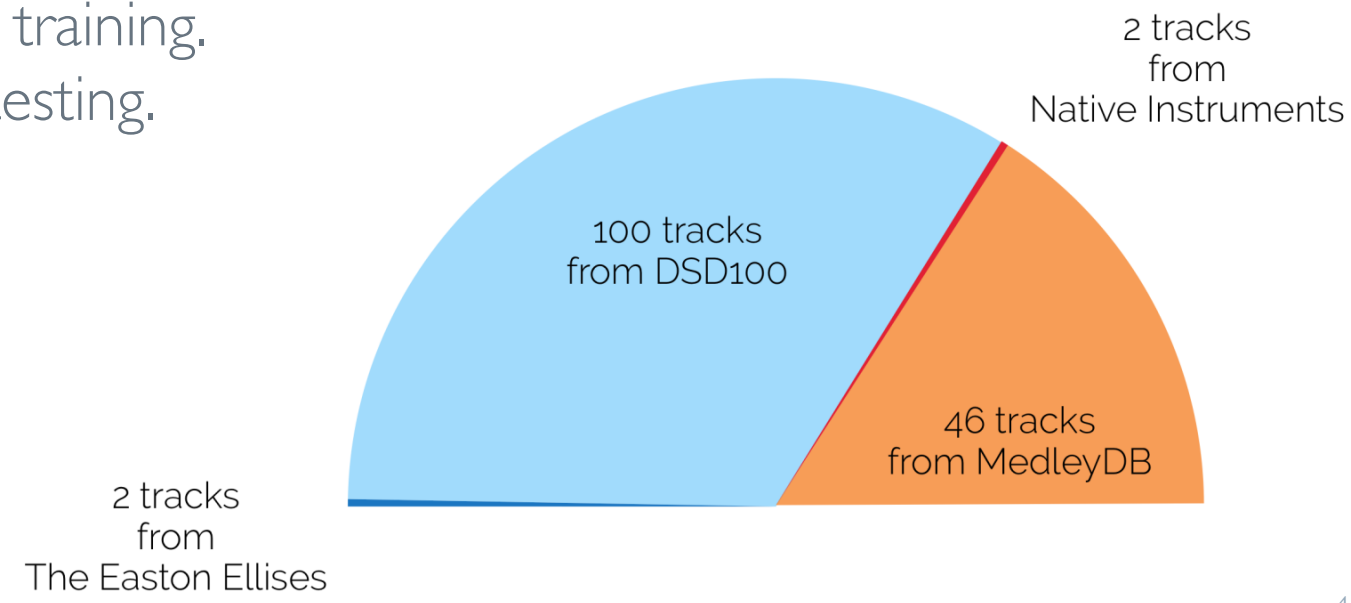
Data augmentation

- Modifying the data with predictable transforms.
- Used in images: rotation, Gaussian noise, crop …

- Used in singing voice processing BUT : simplistic transformations directly on the spectrogram [Schluëter 2016].
- Use case: source separation.
- Need for realistic transformation: separation requires finer details.
- Mix and separated data are rare.

Musdb Dataset

- Musdb dataset :
  - 100 tracks training.
  - 50 tracks testing.



Augmented Dataset

- With augmentation 15,000 tracks ~ 1.5 months.

- Proposed augmentations:
  - Pitch-shifting
    - by a factor p $\in$ {−300;−200;−100;0;100;200;300} (in cents).
  - Time-stretching
    - by a factor t $\in$ {0.5;0.93;1;1.07;1.15}.
  - Transformation of the spectral envelope of the singing voice
    - by a facto $\in$ {−150;−100;0;100;150}.
    - The spectral envelope is estimated and transposed while the pitch remains unchanged.

Specific consideration for each source

Singing voice: pitch shifting using state-of-the-art invariant phase vocoder [Röbel, 2010], performed dynamically on the F0.

Drums: only time stretching.

Bass and accompaniment: transformed using a phase vocoder [Laroche and Dolson, 1999].

Low

Original

High

Experiments

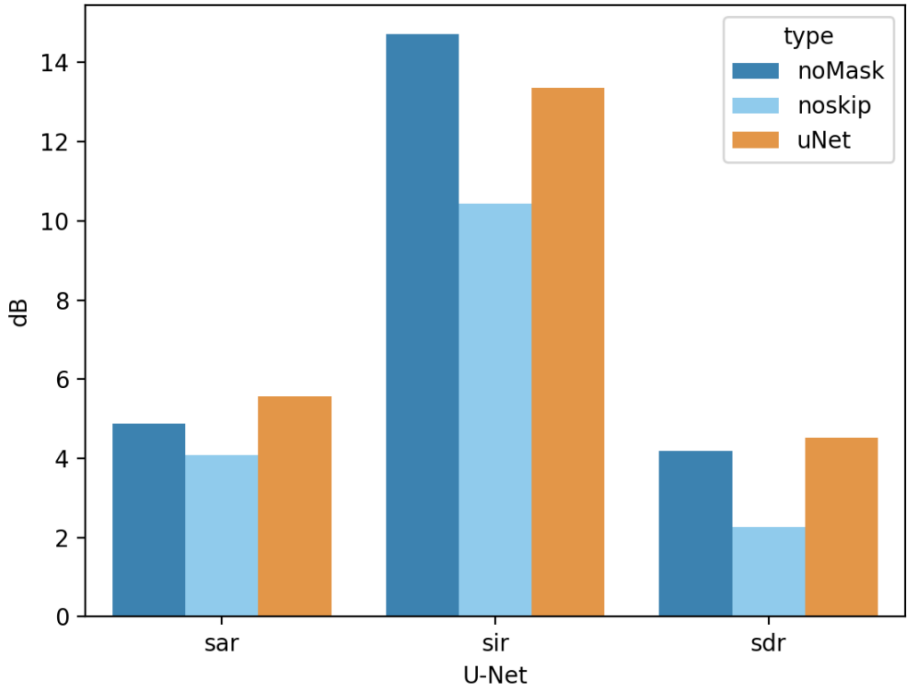
- Study of augmentations
- Architecture of U-Net
  - With or without skip connections
  - With or without masks

Datasets used :

- No-DA : musdb without data augmentation
- DA : musdb augmented, with all combinations.

On U-Net's architecture (side results)

- Skip connections are necessary.
- Outputting a mask helps the results, except for the SIR.



Comparison orignal/adaptated Wave-U-Net

- Same performance between original and adaptated.
- Improvement when using all augmentation.

Augmentation	SDR
No Da	4.09
DA	**4.67**

[Stoller 2018]	3.96

Comparison Augmentations

- Pitch shifting best transformation.
- General improvement for all transformations.

	SAR		SDR	
Augm	U-Net	Wave-U-Net	U-Net	Wave-U-Net
No DA	5.76	5.52	4.52	4.09
Stretch	5.73	5.60	4.85	4.20
Env	6.06	5.23	4.55	3.77
Pitch	**6.35**	**6.09**	**5.20**	**4.67**
DA	6.40	**6.62**	5.30	**5.42**

Results U-net

- Large gap of performance when working on musdb.
- But with augmentation, improvement.
- [Janson et al, 2017] training on 20,000 tracks.
- But does not replace more real data.

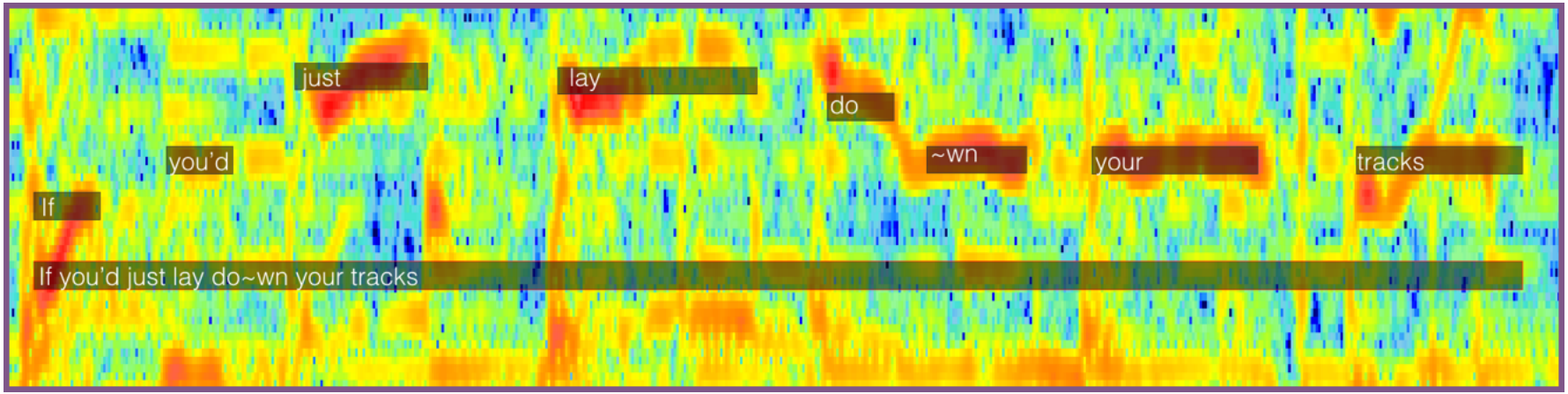
Augmentation	SAR	SIR
No DA	5.76	11.75
DA	**6.40**	**11.98**

[Janson et al, 2017]	11.30	15.31

Creating DALI: online ressources and Teacher/Student

Creation of a dataset

- Goal: dataset of audio aligned with meldoy notes and lyrics.
- What for ? audio to lyrics alignement, lyrics translation, singing voice detection, F0 estimation.



Karaoke resources

```
#ARTIST:Dire Straits
#TITLE:Money For Nothing
#MP3:Dire Straits - Money For Nothing.mp3
#EDITION:80s
#GENRE:Rock
#LANGUAGE:Englisch
#BPM:268
#GAP:530
#VIDEO:Dire Straits - Money For Nothing.AVI
: 16 10 22 I
: 30 8 21 want
: 42 10 17 my
* 56 8 14 M.
* 68 12 19 T.
* 82 60 21 V.
- 160
: 1114 2 5 Now
: 1118 2 10 look
: 1121 2 10 at
: 1124 4 10 them
: 1130 2 7 yo-
: 1134 2 5 yo'
: 1138 2 2 s
- 1144
```

Time onset

Text

Duration

Musical note

Karaoke file

How to find the correct audio ?

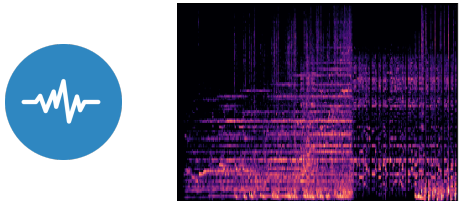
- With the name and artist in the karaoke file, request to Youtube.
- But several versions of one track:
  - Radio edit
  - Live ….

Problems are :

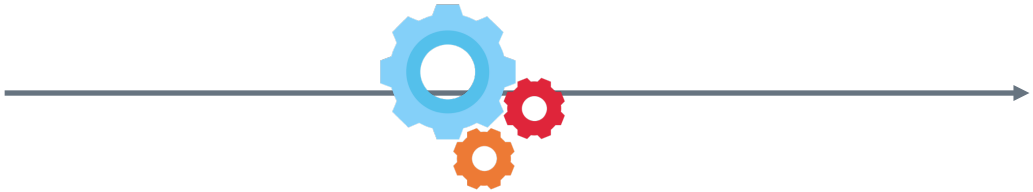
1. How to find the correct ones ?
2. Are annotations good enough ?
3. Do annotations need adaptations ?

Need for a common representation between karaoke file and audio.

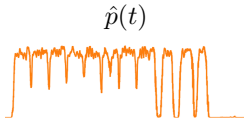
Comparing audio and karaoke files (1)



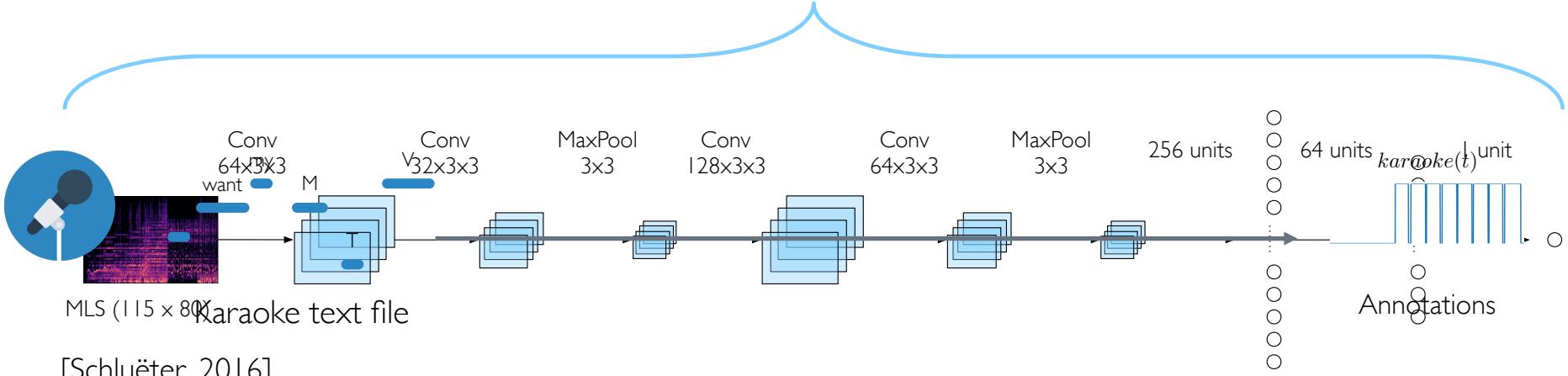
Audio track



Singing voice detection system

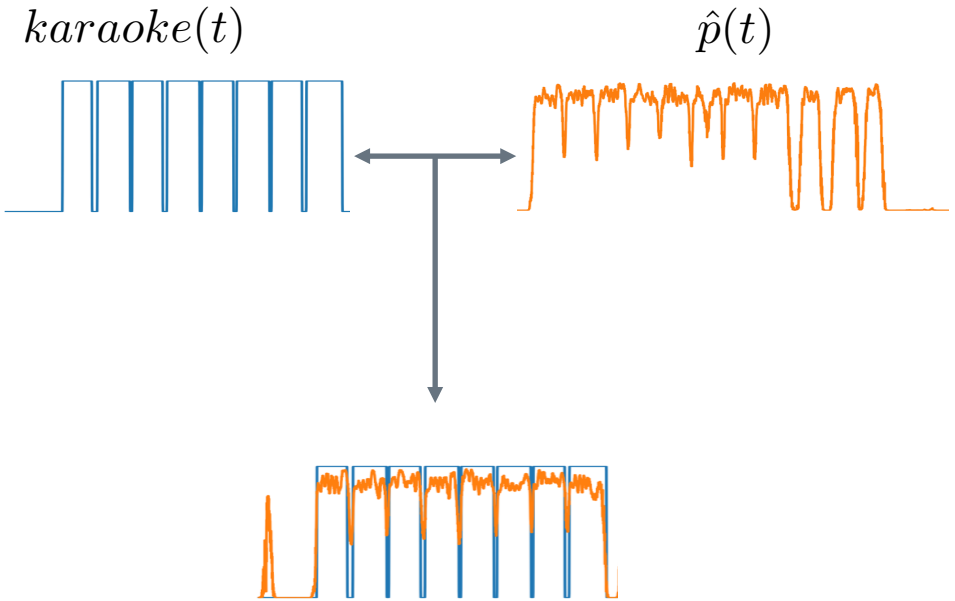


Singing voice probability



[Schluëter, 2016]

Comparing audio and karaoke files (II)



Normalized cross-correlation (NCC) as distance:

$$NCC(o, fr) = \frac{\sum_t karaoke_{fr}(t-o)\hat{p}(t)}{\sqrt{\sum_t karaoke_{fr}(t)^2}\sqrt{\sum_t \hat{p}(t)^2}}$$

The NCC gives us a score of how well the track is aligned.

Tracks selection

With the NCC score, selection of the tracks that have a high alignment score.

Improvement on our singing voice detection system

Our alignement between a karaoke file and the audio is only as good as our singing voice detection system.

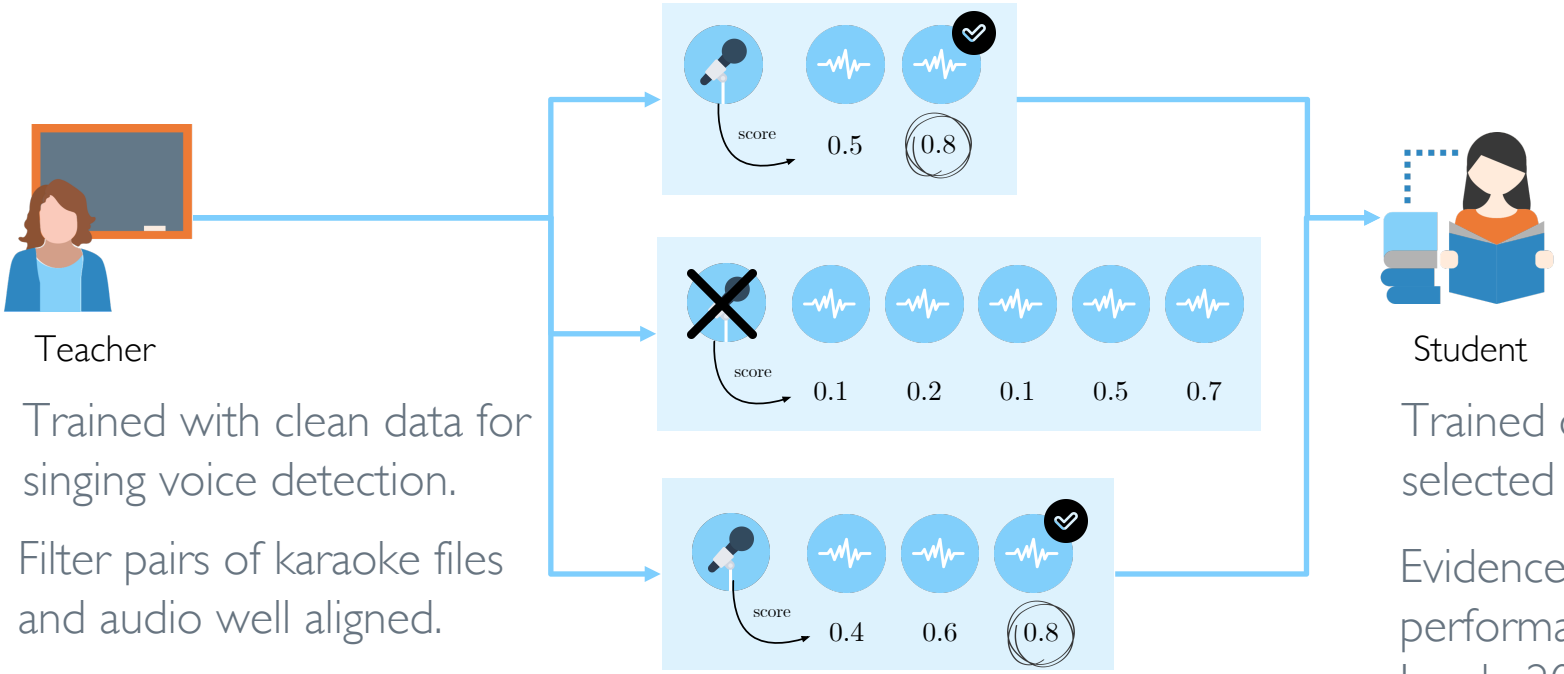
With the score of the NCC : find **which tracks are aligned best**.

13,339 karaoke files to **2,440** (audio, karaoke files) "well aligned" (NCC score > 0.8).

To re-use this data :

Teacher/Student paradigm

Teacher/ Student paradigm



Teacher

Trained with clean data for singing voice detection.

Filter pairs of karaoke files and audio well aligned.

Student

Trained on the tracks selected by its teacher.

Evidence of better performances [Wu & Lerch, 2017].

Experiments

Datasets:

- Use of 2 singing voice dataset.
  - Jamendo (93 tracks, 61 for training).
  - Medley DB (122 tracks).
- We also use a fusion of MedleyDB + Jamendo.

Experiments:

- One teacher per dataset (3 teachers total).
- One student per teacher.
- Each teacher selects some tracks for its student.

Results

- Cross datasets results.
- Students generally outperform teachers.
- Use best student (Student J+M) for final filtering.

SVD System		Jamendo test set	Medley DB test set
Jamendo	Teacher	87%	**82%**
	Student	82%	**82%**
Medley	Teacher	**76%**	**85%**
	Student	**80%**	**84%**
J+M	Teacher	**82%**	**82%**
	Student	**86%**	**87%**

At the end DALI Version 1

Files of DALI:

```
"lines": [
    {
      "text": "go tell it on the mountain",
      "freq": [
        440,
        698.4564628660078
      ],
      "time": [
        2.5719999999999996,
        6.696999999999999
      ],
      "index": 0
    },
```

- 3 levels of annotations: notes, lines, paragraph
  - Notes
  - Lyrics

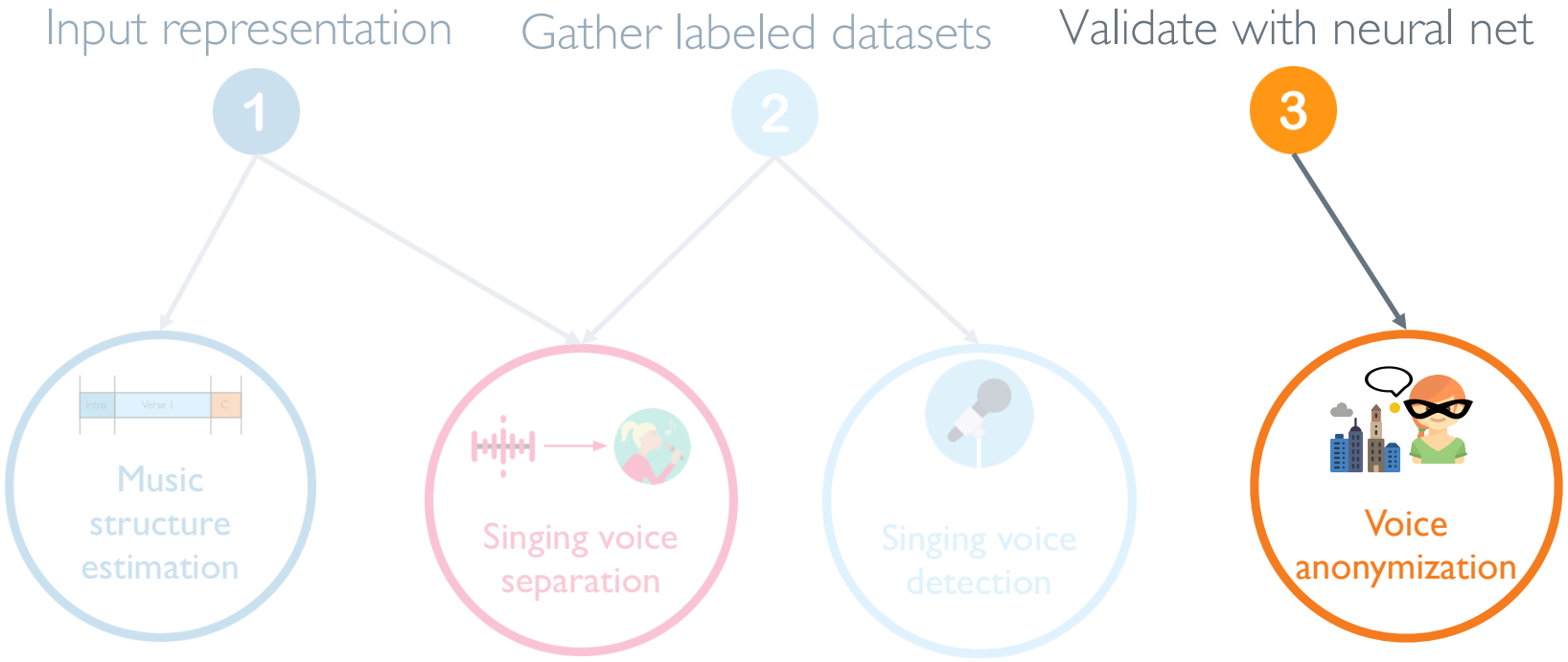
Version	Songs	Artists	Mean songs per artist	Top 3 genres	Top 3 languages	Top 3 decade
DALI 1.0	5,358	2,274	2.3562	Pop: 2,662 Rock: 2,078 Alternative: 869	English: 4,018 German: 434 French: 213	2000s: 2,318 1990s: 1,020 2010s: 6,68

Partial conclusion - datasets

- Having more data is beneficial.
- Two ways to have more data:

  - Traditionnal: Data augmentation.
    - Does not replace new data.
    - Useful if no ressources online.
  - New method: Teacher/Student.
    - Use of online ressources.
    - Improvement using new learning paradigm.

To study those problems: 4 tasks



IV.

How to use neural networks to validate solutions?

Use case: Voice anonymization

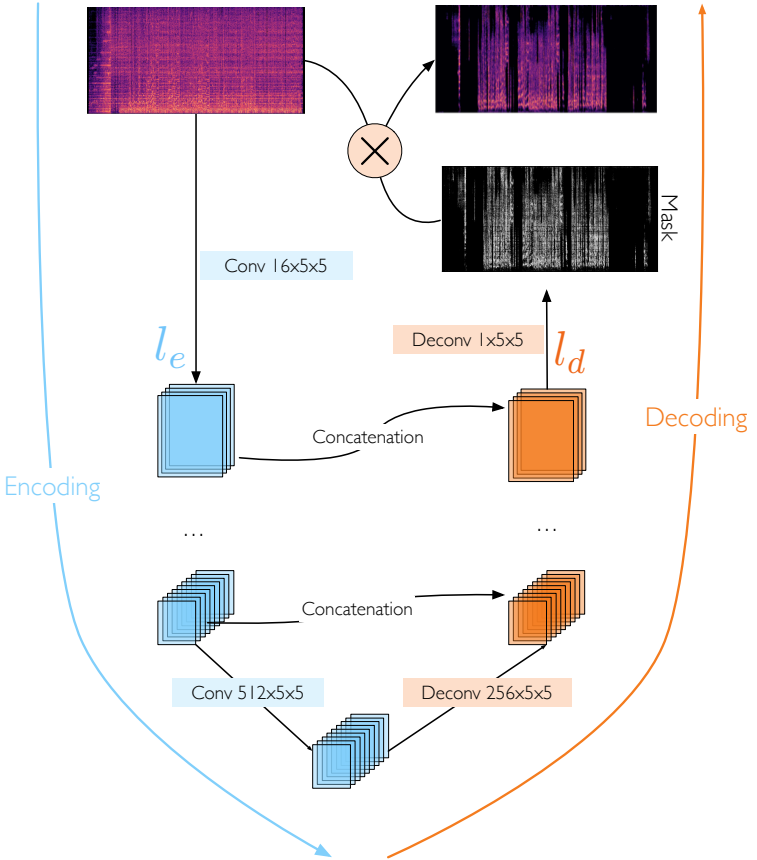
Work published in [Cohen-Hadria et. al, 2019b]

Introduction

- Smart cities see the emergence of large sensor networks, monitoring useful data for the city's running.
- In sound recording networks: sometimes, pick up human conversations.
- Need for an anoymizaton method of those recordings.
- 3 criteria:

First step - Separation

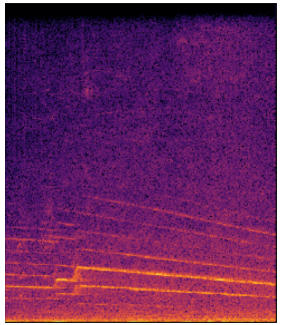
- To preserve the scene: first step of separating the voice from the background.
- To extract the voice: U-Net model.



Creation of the dataset

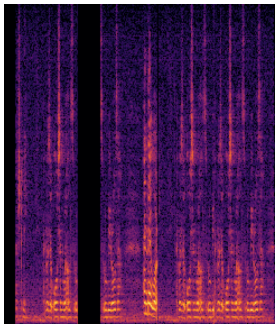
- Synthetic dataset.
- Background from SONYC-UST.
- Voices from VoxCeleb.

Backround



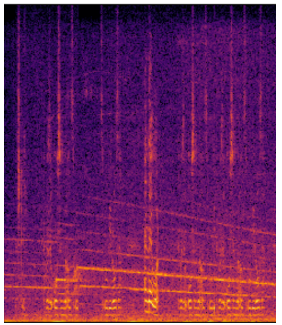
+

Voice



=

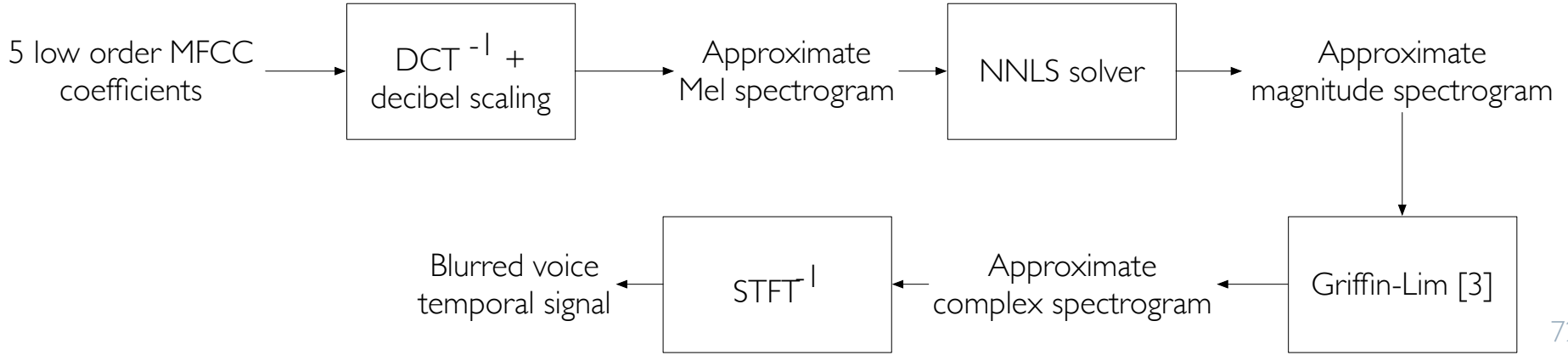
Mix



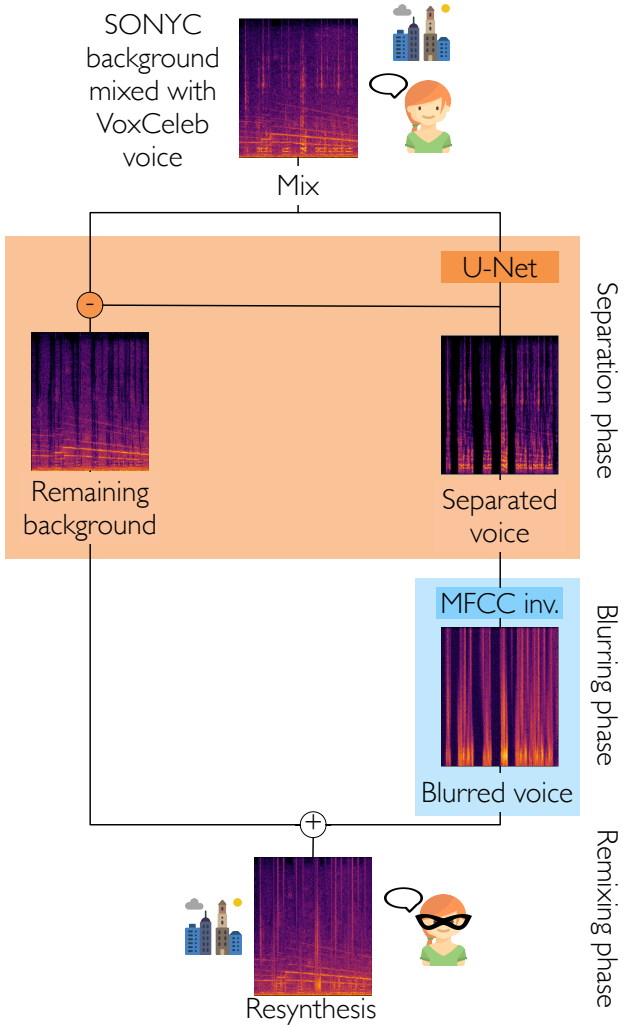
Blurring

After the separation step: blurring on the separated voices. We propose two blurring methods:

1. Blurring with a low pass filter at 500hz.
2. Blurring with MFCC inversion.



Complete method of blurring



Experiments

- Both automatic and human evaluation.
- Two voice to background ratios: **High** and **Low.**
- Designed to assess our three goals:

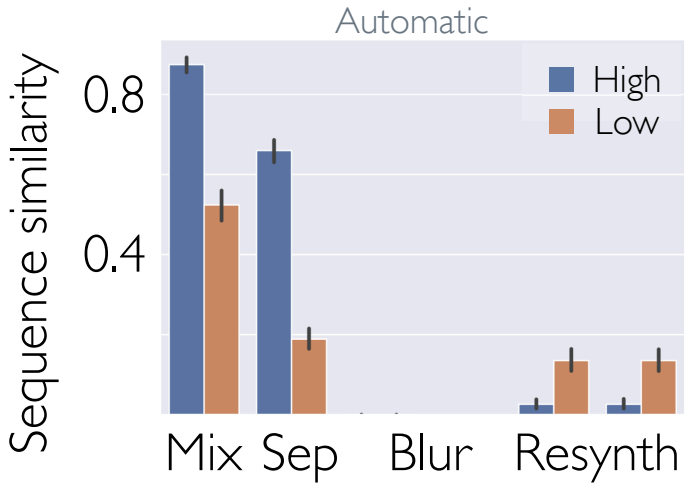
Experiment 1: Content obsfucation

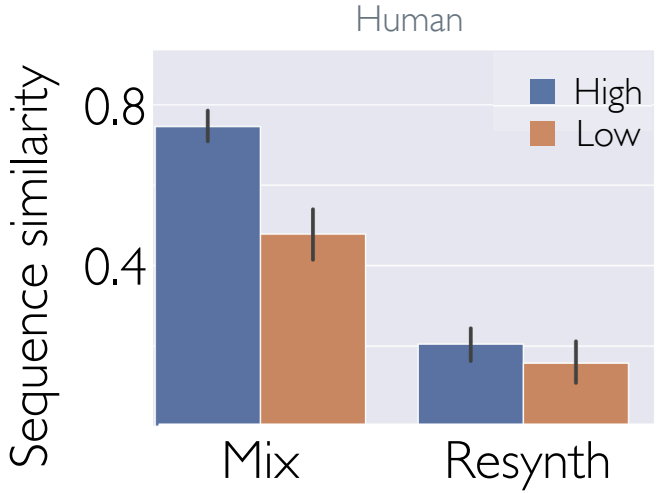
- **Dataset**: LibriSpeech dataset (contains transcript of voices).
- **Automatic evaluation**: Automatic Speech Recognition (ASR) system (Google API).
- **Human evaluation**: transcribe what you understand.
- **Metric:** Sequence similarity.

Experiment 1: Content obsfucation

1 Obfuscate the content

- Lower is better (masking content).
- Blurred versions are never transcripted.
- Only resynthesis does not fully obfuscate the content -> due to the quality of the separation.
- Experiment validated by humans.





Experiment 2: Masking identity

- **Dataset**: Use of VoxCeleb, containing recordings of celebrity.
- **Automatic evaluation**: VggVox model for speaker identification.
- **Human evaluation**: hard to do.
- **Metric**: % correct indentification.

Experiment 2: Masking identity

2

Mask the speaker identity

- Lower is better (masking identity).
- For both high and low, our blurring method decreases the identification.
- Need for human evaluation, but necessitates training.

SNR	Audio	% correct id
High	Mix	83
	Low pass filter	43
	MFCC inversion	43
Low	Mix	43
	Low pass filter	29
	MFCC inversion	29

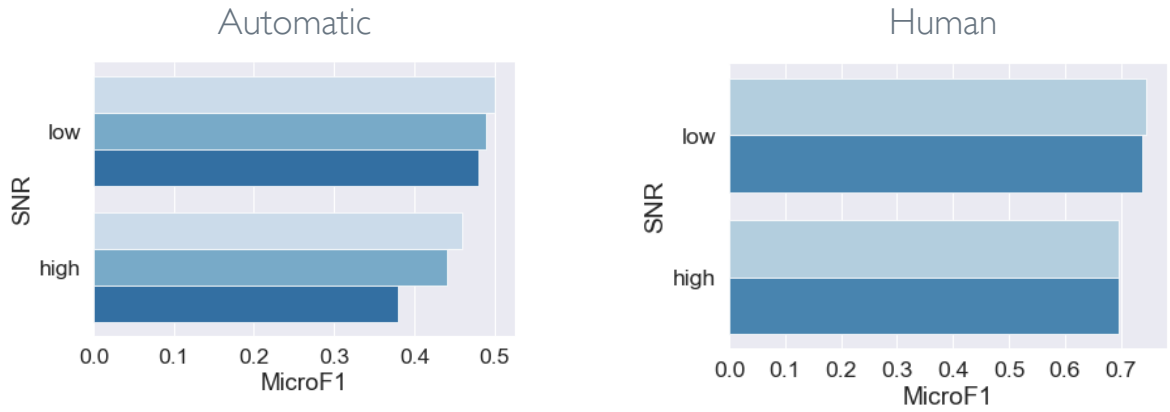
Experiment 3: Scene preservation

- **Dataset**: Use of SONYC-UST dataset (+ voices from VoxCeleb). Contains labels for 8 coarse classes.
- **Automatic evaluation**: DCASE 2019 baseline for urban sound tagging.
- **Human evaluation**: What do you here in these scenes?
- **Metric**: classification F1 score.

Experiment 3: Scene preservation

3 Preserve the acoustic scene

- No differences on the classification results (preserve scene).
- Our blurring method preserves the acoustic scene.
- Confirmed by human experiments.



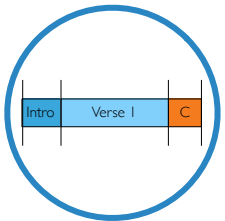
Demo

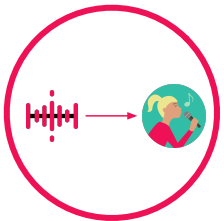
Partial conclusion

- Proposed original method of blurring voice.
- Hard to evaluate by humans (speaker ID).
- Proposed automatic evaluation using:
  - ASR system.
  - ConvNets for speaker ID and scene classification.

- When comparing to human evaluations:
  - Validate automatic evaluation with neural networks.

V.
Conclusion



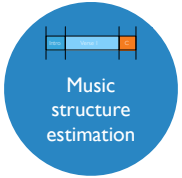

Conclusion

1. About input representation:
   - Using prior knowledge helped for structure estimation.
   - For singing voice separation: if more data waveform. Other case, spectrogram.
2. Gathering of large datasets:
   - Data augmentation: useful when no other resources. Does not replace real data.
   - Teacher/Student paradigm: helped align data and create automatically real data.
3. Validate with ConvNet:
   - Blurring techniques evaluated automatically.
   - Automatic and human evaluation correspond.

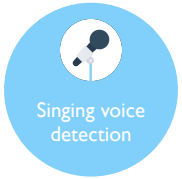
Future works

1. About input representation:
   - Study different types of fusion (early, late).
   - Input depth: what to stack?
   - Other information (singing voice, beat).
2. Gathering of large datasets:
   - Data augmentation: comparative study of augmentation.
   - Student/teacher paradigm: 2nd generation.
   - Student/teacher paradigm: different roles for teacher.
3. Validate with ConvNet:
   - Human evaluation on speaker id.

Thank you for your attention!



Cohen-Hadria, A. and Peeters, G. (2017).
Music Structure Boundaries Estimation Using Multiple Self-Similarity Matrices as Input Depth of Convolutional Neural Networks.
In *AES International Conference Semantic Audio 2017, Erlangen, Germany*



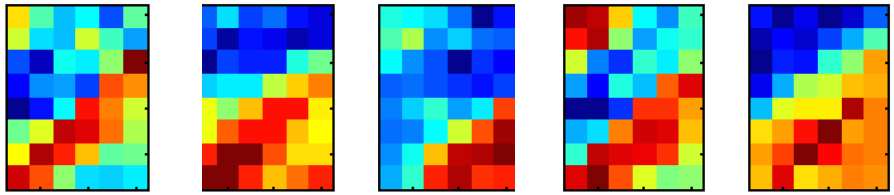
Meseguer-Brocal, G., Cohen-Hadria, A., and Peeters, G. (2018).
Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher student machine learning paradigm.
In *Proceedings of International Society for Music Information Retrieval Conference (ISMIR), Paris, France*

Cohen-Hadria, A., Roebel, A., and Peeters, G. (2019a).
Improving singing voice separation using deep U-Net and Wave-U-Net with data augmentation.
In *2019 28th European Signal Processing Conference (EUSIPCO), La Corona, Spain*

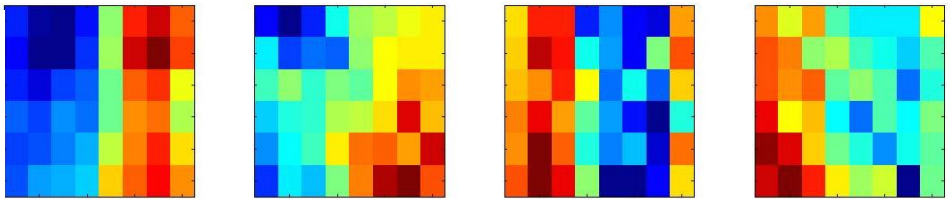
Cohen-Hadria, A., Cartwright, M., McFee, B., and Bello, J. P. (2019b).
Voice anonymization in urban sound recordings.
In *29th IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Pittsburgh, Pennsylvania, USA*

Filters learned on different representations

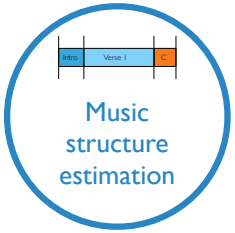
SSM



Spectrogram



Future works



Study different types of fusion.

Other information (singing voice, beat)

Listening test

Comparative study of augmentation

2nd generation

Study on alignment

Different role of teacher

Human evaluation on speaker id

More control on blurring techniques